30749

NATIONAL LIBRARY
OTTAWA

CANADA

BIBLIOTHÈQUE NATIONALE
OTTAWA

NAME OF AUTHOR.....Wo-Shun Luk.....

TITLE OF THESIS...Analytic Evaluation of...
...Information Retrieval Processes...
.....

UNIVERSITY...of...Alberta.....

DEGREE FOR WHICH THESIS WAS PRESENTED.....Ph.D.....

YEAR THIS DEGREE GRANTED.....1976.....

Permission is hereby granted to THE NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed).....

PERMANENT ADDRESS:
#1104 - 8375 - 112 Street,
Edmonton, Alberta
.....

DATED..Aug. 13.....19 76

NL-91 (10-68)

INFORMATION TO USERS

THIS DISSERTATION HAS BEEN
MICROFILMED EXACTLY AS RECEIVED

This copy was produced from a microfiche copy of the original document. The quality of the copy is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Canadian Theses Division
Cataloguing Branch
National Library of Canada
Ottawa, Canada K1A ON4

AVIS AUX USAGERS

LA THESE A ETE MICROFILMEE
TELLE QUE NOUS L'AVONS RECUE

Cette copie a été faite à partir d'une microfiche du document original. La qualité de la copie dépend grandement de la qualité de la thèse soumise pour le microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

NOTA BENE: La qualité d'impression de certaines pages peut laisser à désirer. Microfilmee telle que nous l'avons reçue.

Division des thèses canadiennes
Direction du catalogage
Bibliothèque nationale du Canada
Ottawa, Canada K1A ON4

THE UNIVERSITY OF ALBERTA

ANALYTIC EVALUATION OF INFORMATION RETRIEVAL PROCESSES

by

© WO-SHUN LUK

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

IN

COMPUTING SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

FALL, 1976

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the faculty of Graduate Studies and Research, for acceptance, a thesis entitled ANALYTIC EVALUATION OF INFORMATION PROCESSES submitted by WO-SHUN LUK in partial fulfilment of the requirements for the degree of Doctor of Philosophy

.....
Supervisor

.....

.....

.....

.....
External Examiner

Date: August 12, 1976

TO MY PARENTS

ABSTRACT

Two processes, i.e., relevance feedback and retrieval on clustered files, are modelled and analysed. Experiments are also conducted to verify part of the theoretical results. For each individual process, the behavior of the system performance is studied under the variation of the key parameters of the process. Together, the processes lend themselves as examples for studying modelling and analytic techniques for evaluating information retrieval processes.

ACKNOWLEDGEMENT

I would like to thank my thesis supervisor, Dr. C. T. Yu for his guidances and continued support. He has also initiated the research topics covered by this thesis. Thanks are due to members of my supervisory committee and my external examiner, Dr. G. Salton for their suggestions and criticisms.

I am grateful to my wife, Lydia, for her patience and understanding during the years of my graduate study, as well as her invaluable assistance in typing this thesis.

TABLE OF CONTENTS

		PAGE
CHAPTER 1		
1.1	Problem Area	1
1.2	Information Systems	2
1.3	System Effectiveness and Efficiency	5
1.4	System Evaluation	8
CHAPTER 2		
2.1	Motivation	10
2.2	A Probabilistic Model	14
2.3	Some Necessary and Sufficient Conditions	21
2.4	Optimal Values for α and β	29
2.5	Experimental Results	32
2.6	Generalization	39
CHAPTER 3		
3.1	Introduction	41
3.2	Pelated Work	42
3.3	A Probabilistic Model	45
3.4	Analysis	50
3.5	Empirical Results	61
3.6	Conclusion	64
CHAPTER 4		
4.1	Distribution of Similarities	66
4.2	Choice of Model	68
4.3	Mathematical Manipulation	70
4.4	Summary	72

- cont'd -

PAGE

REFERENCES 73

TABLES 1 - 5 76

APPENDIX I 81

APPENDIX II 83

APPENDIX III 88

CHAPTER 1

INTRODUCTION

1.1 Problem Area

This thesis addresses itself to the problem of the analysis of processes in information retrieval. Two important processes, namely, relevance feedback (RF) and retrieval in clustered files (RCF) are selected as candidates for detailed investigation. The purpose is actually two-fold. As a practical, short-range goal, the analysis will reveal the intrinsic relationships among various key parameters of the processes, indicate regions in the parameter space which guarantee good results, and, in some cases, derive optimal values for the parameters. These analytical results should prove useful to those designers of information systems who wish to adopt these processes. On the other hand, the modelling, as well as the analytic techniques will hopefully serve as valuable examples to others with similar research interest. Two different models are constructed for the processes. The model used for RF is developed from Swets' continuous model, in which the items and attributes are invisible. It is by and large a "macroscopic" model. For RCF, however, a discrete and "microscopic" model is used which heavily depends on the occurrences of each attribute. Chapter 4 will be devoted to further explaining details of the

two models and contrasting one to another.

1.2 Information Systems

As a branch subject of computer science, information retrieval is not very well defined; in fact, a large part of non numeric computing activities can be classified as information storing and retrieving. It is therefore appropiate to first define, before proceeding with the main body of this thesis, the kind of information systems on which the subsequent discussions are based.

The major components of an information retrieval system are depicted in Fig. 1. Contained in the data base are a set of records, each of which represents an item in the "information base", where the ultimate information needed by the users is stored. The basic unit of an information base is an item, which can be a document, a personnel record, a description of an auto part or an antique in the museum etc. Corresponding to each item there is a record in the data base which consists of a set of attributes, chosen to represent the item. Through the process of indexing, the information base is converted into a data base which is structured for computer searching.

To describe the functioning of the system, we start with the user who requests information. This request is often expressed as a query which, like a record in the data base, also consists of a number of attributes. The query is

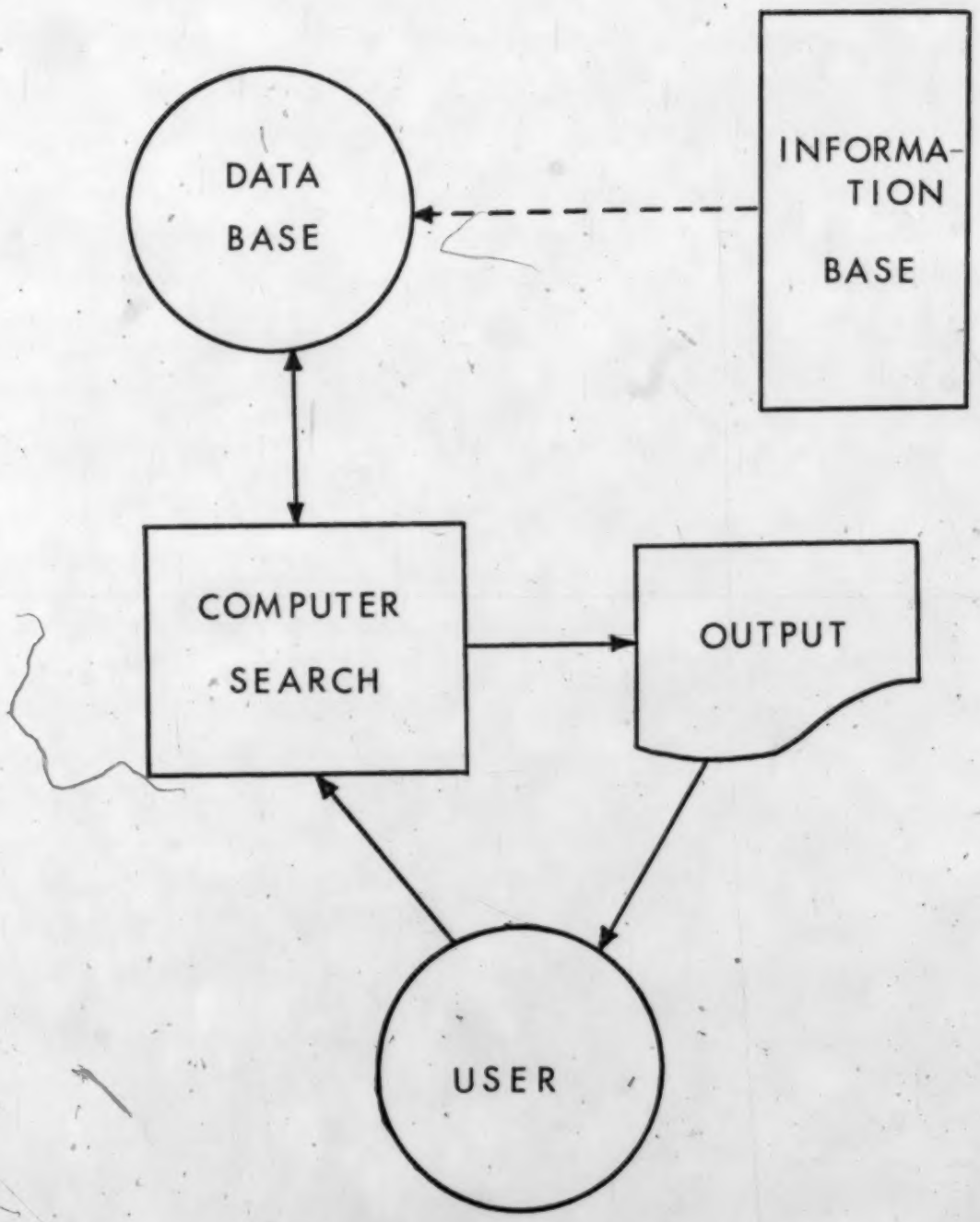

Fig. 1

then submitted to the system which matches it with each record in the whole data base (or its subset). In the type of system discussed here, best-match method is used and only those which are considered as "close" to the request will be retrieved (the concept of closeness will be clearly defined later). The number of items retrieved is usually controlled by an input parameter, called the threshold, which is determined either by the user or the system manager. The system is an on-line system. The user can interact with the system through some communication channel, like a terminal.

To help analyse the whole retrieval process, the terms introduced above have to be more rigorously defined. Let n be the total number of attributes in the system. Each record in the data base is then generalized as a n-tuple where the ith component represents the values for the ith attribute. A query is similarly defined. A value of 0 in the ith component indicates that the attribute is not related to the item represented by this record. The higher the value assigned to the ith component, the more important the ith attribute is considered to be to the item. In some cases however, a record (and query) might be simplified as a binary vector. The similarity between an item and a request is quantified by the similarity between the record and the query representing them. Here we express the similarity between two n-vectors $A=(a_1,\ldots,a_n)$ and $B=(b_1,\ldots,b_n)$ by means of the simple matching function

$$f(A,B) = \sum_i a_i b_i \qquad (1.1)$$

If A and B represent a query and a record respectively, a larger function value means that the record is closer to a query and hence that the record has a better chance of being retrieved. (In particular, if all the records and queries in the system are binary vectors, f will take as its value the number of attributes in common between the query and the record.) The record will be retrieved if and only if $f(A,B) > t$, where t is a pre-assigned threshold value. Alternatively, the user can restrict the number of retrieved items, say 10, so that only the 10 items with the highest function values are retrieved.

Geometrically, the items and queries can be regarded as points (or vectors) in the n-dimensional space, the distance between the vectors being measured in some norm. With the threshold t as the radius and the query as the centre, all the items that fall within this sphere are retrieved. Alternatively, if only 10 retrieved items are required, they will be the 10 closest neighbors of the query.

1.3 System Effectiveness and Efficiency

The best-match type of retrieval is not necessarily applicable to all information systems. Opposed to the idea of best-match is the exact-match which retrieves all those, and only those records matching exactly with the query, i.e.,

containing all the attributes of the query. So, in those systems using the exact-match method,all the items retrieved will be pertinent to the user's needs. Apparently, this matching method does not always satisfy all users. For instance,in a library environment, the user of the system usually has a very vague idea of what he actually needs. He might want to find references on some subject, but does not know the authors or titles. He can only roughly describe the contents of the documents or books he needs by a few keywords, the choice of which is obviously a subjective one. It is therefore desirable to let the system determine which documents are most likely to be useful to him. Some advanced information systems like MEDLAR (by American Medical Library Association) and SMART (an experimental system for the researchers at Cornell University) have adopted such a method.

Together with retrieval by content comes the problem of relevance. For a system to be 100% effective, all the items that are considered as relevant to the request must be retrieved and every item retrieved must be relevant. This ideal situation is rarely achieved simply because the ultimate judgment on whether an item retrieved is useful or not, is made by the user who initiates the request. This problem persists to some extent even in a completely manual information system.

On the other hand, computerized information systems,

thanks to their earlier successes, are gaining in popularity. Coupled with the fact that the cost of building and using one is decreasing, the users are demanding ever larger systems and their application is expanding into new areas. To meet these challenges, new processes have been devised, aiming at improving the system effectiveness and/or efficiency. Here are a few examples of such processes.

(1) Manual indexing can no longer cope with the explosive information growth. Researchers are now looking into automatic text-processing methods, which will undoubtedly improve the efficiency.

(2) If the data base is organized into different classes according to their contents, it becomes possible to search selectively some parts of the data base, thus saving a lot of time.

(3) It can be beneficial to the user to communicate with the system his relevance decision on the items retrieved so that the system can utilize this feedback of information to retrieve more items that may be useful to him. In this case, more computing time as well as the user's time will be consumed, but the user will probably be more satisfied with the retrieval result.

However, implementation of each of these processes requires a tremendous amount of both human effort and financial resources. There has to be some means of evaluating

these processes to see whether they are justified. For example, can automatic indexing compete with manual methods in term of producing retrievals of equally high quality? What is the risk of deteriorating retrieval performance by ignoring other parts of the data base? Can the feedback method really improve the system effectiveness? Procedures should be established to provide answers to questions like these.

1.4 System Evaluation

The first systematic approach to system evaluation was adopted by the famous Aslib project in Cranfield, England, in early 1960's. There, experiments were conducted to examine, among other things, different indexing strategies. The sample collections employed were documents on aerodynamics and aircraft structure. The size of the collections ranged from 82 to 1400 documents per collection. The sample queries were submitted by aeronauticists and each document in the collection was also manually examined to determine its relevance to the query. The effectiveness of an indexing strategy was measured by recall R and precision P, defined as:

$$R = \frac{\text{number of items retrieved and relevant}}{\text{total relevant in collection}},$$

and

$$P = \frac{\text{number of items retrieved and relevant}}{\text{total retrieved in collection}}.$$

The SMART system [Salton 1968] greatly enhanced this method and automated it. More sample data bases on different subjects were added to the Cranfield collections and the system was capable of testing many more processes. The system is now available as a software package as testing ground for system designers and researchers to evaluate their newly devised methods as well as various input parameters to the already known processes. The work carried out by the information specialists at Arthur D. Little, Inc [Giuliano 1966] is similar in nature. This approach is still being used extensively and is generally considered acceptable by the industry. Nevertheless, with so many input parameters usually associated with each process, there are no assurances that the values chosen will in fact be optimal or near optimal (in some sense), or indeed will work at all. Recently, research articles that are rather theoretical in nature have emerged in this area [Brookes 1968, Swets 1969, Bookstein 1974, Yu 1975 etc.]. However, most of them are mainly concerned with building the models rather than making use of them for some specific processes. Others emphasize indexing strategies. This thesis is the first real attempt to bring these well-known processes (i.e., RF and RCF) and the models together. In doing so, not only are the models put to use for more constructive purpose, but also the models are tested for their shortcomings and adequacy for mathematical manipulation.

CHAPTER 2

RELEVANCE FEEDBACK

2.1 Motivation

It is generally conceded that there is plenty of room for new techniques that aim at improving the effectiveness of a computerized retrieval system. The computer is not an intelligent machine (no breakthrough is yet in sight in the efforts of making it one) and man-machine communication is far from being perfect. Two possible remedying strategies to increase the interaction between the user and the system can be adopted. The data base can be "tuned" regularly based on the users' response on the previous retrieval performance. This involves changing the index representation of the data base. Quite a number of methods have been proposed and analyses of selective methods are provided [Yu 1976]. Alternatively, the user's query can be altered by the system in an interactive environment. The user evaluates each of the retrieved items as either relevant or irrelevant and then sends the information back to the system. The system then formulates a new query by making use of such information. Hopefully, this new query will retrieve more relevant items and fewer irrelevant items. This process is called relevance feedback. This process has been designed mainly for an on-line document retrieval system, where the

items are actually documents and the users are searching for quick references on some specific topics. Here, we are concerned more with the relevance of the retrieved documents than the representation of the documents in the system. Therefore, in the rest of this chapter, we shall indiscriminately refer to an item or the record representing it as a document.

A practical method for updating queries has been suggested by Rocchio [1965]. The new query is given by

$$Q^* = Q + \alpha\sum D - \beta\sum D' \qquad (2.1)$$

where $\alpha, \beta \geq 0$ are parameters, and D and D' sum over respectively the sets of relevant (R) and irrelevant (I) items retrieved by Q. The formula has the effect of increasing the influence of relevant documents (and hence, the attributes contained in them) and decreasing the effect of the irrelevant ones. There have been a lot of experiments conducted which show that this particular algorithm performs reasonably well [Rocchio 1965, 1966]. The behavior of some variants of the method, such as deleting one of three terms in the equation has also been observed [Ide 1968, Crawford 1968]. However, very little theoretical justification has been given.

Intuitively, the relevance feedback method should improve the retrieval performance since the system has obtained more information from the user about his require-

ments. But it might fail if the relevant items are too dispersed or the user's query is too ill-formulated (this will happen if the user is too vague about what he actually wants).

To explain this phenomenon in greater detail, let us consider a hypothetical system with only two attributes. In this way, each document can be adequately represented by a point in a plane (see Fig.2). Suppose that the system retrieves 5 documents in the first try, 2 relevant and 3 irrelevant. All the 2 relevant documents retrieved are located in the top-left of the retrieval circle while the 2 irrelevant ones are in the opposite position. Under the effect of (2.1), the query will be shifted in the direction of the relevant ones and away from the irrelevant ones. If the relevant documents (and to a lesser extent, the irrelevant ones) are "flocked" together (as in Fig.2(a))the new query thus generated will produce better retrieval.† However, if the relevant documents are dispersed (see Fig.2(b))

---

†It is interesting to note that when there are two or more "flocks"of relevant documents retrieved, the query could as well be split up into a number of new queries, maybe one for each such "flock". There are usually relatively few relevant documents retrieved each time (that is why the feedback method is needed!), so it would be difficult to detect multiple flocks. Neverthesless experiments [Borodin 1968] have been conducted to test the split query method, although this method will not be dealt with here.

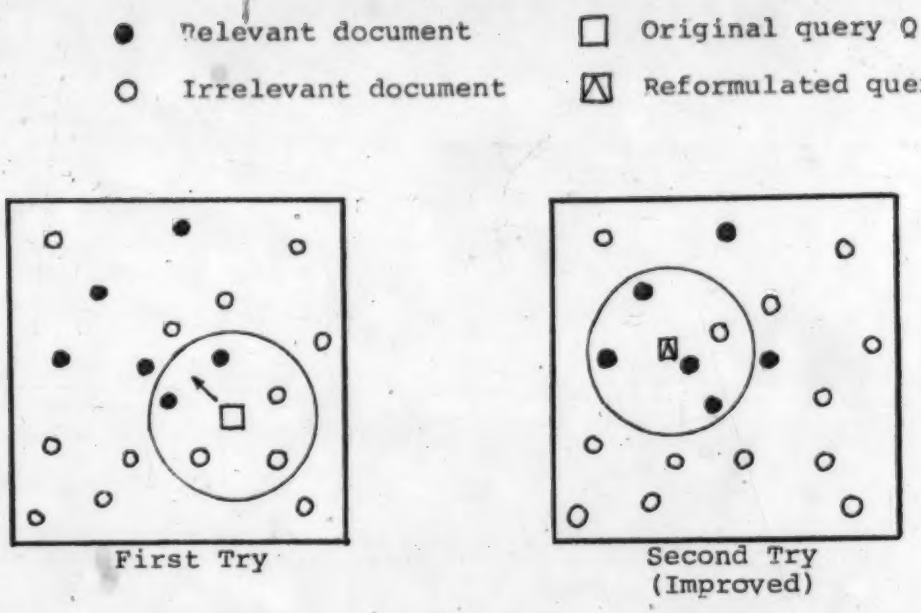

Fig. 2(a) "flocking" of relevant documents

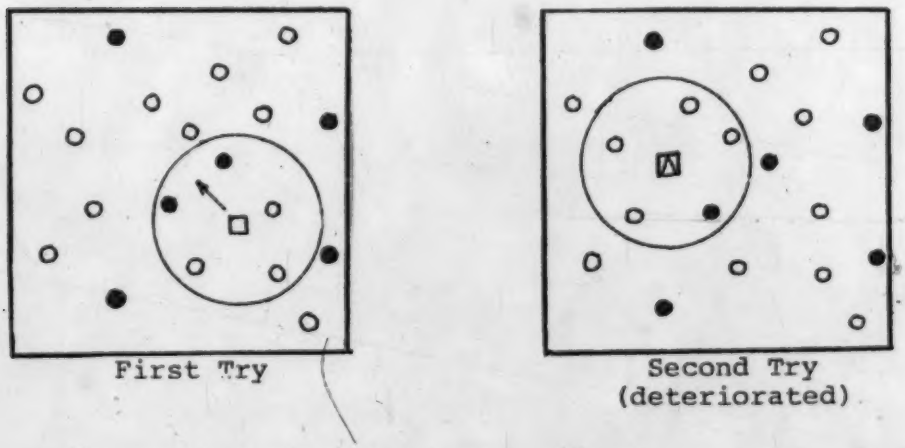

Fig. 2(b) dispersion of relevant documents

the retrieval performance can deteriorate.

The other possible cause of failure of the method is that the system cannot distinguish the relevant documents

from the irrelevant ones. It occurs when the query is not adequately prepared in the first place, such as containing attributes with opposing meanings etc., or the data base is not organized properly with respect to this particular query in question. As a result, the set of relevant documents as a whole is no longer "closer" to the query than the set of irrelevant documents.

It is these arguments that motivated the analysis of relevance feedback. The arguments clearly indicate that some concepts have to be quantified in order that rigorous analysis can be carried out. Among them are the idea of documents "flocking together" and the distinction between relevant and irrelevant documents. In the next section, a probabilistic model will be constructed which will enable these concepts to be precisely defined. The arguments also pave the way along which the analysis can develop. In fact, the analysis has succeeded in verifying these intuitive arguments.

2.2 A probabilistic Model

With respect to a query Q submitted by the user, the document space is divided (at least implicitly in the user's mind) into two disjoint subsets, namely the set $\bar{R}$ of relevant documents and the set $\bar{I}$ of irrelevant documents. Obviously, $R \subset \bar{R}$ and $I \subset \bar{I}$, where R and I are defined in Section 2.1. In the following, for each given query Q, we shall

define six classes of normal distributions over $\bar{R}$ and $\bar{I}$.

The first normal distribution is for the random variable which is the inner product $f(Q,D)$ between Q and a document D in the population set $\bar{R}$. The normal distribution is the relative frequency of documents D which assume the value $f(Q,D)$. The expected value and standard deviation of this distribution are assumed to be μ_1 and σ_1 respectively. Similarly, we define the normal distribution for the variable $f(Q,D')$ over the population set $\bar{I}$ with expected value μ_2 and standard deviation σ_2. (The distributions presented so far are those of Swets [Swets 1963, Brookes 1968]). Next, for each retrieved relevant document $D_1 \epsilon R$, we can define a normal distribution for the variable $f(D_1,D_2)$, $D_2 \epsilon \bar{R}$. We assume that all of these distributions have the same expected value μ_3 and standard deviation σ_3 (i.e. they do not depend on the individual D_1)[†]. Similarly, we define the other three classes of normal distributions for $f(D_1,D_2')$, $f(D_1',D_2)$ and $f(D_1',D_2')$. These distributions are summarised in the table below.

[†]For later development, it is sufficient that the random variable $\frac{1}{|R|} \sum_{D_1 \epsilon R} f(D_1,D_2)$, $D_2 \epsilon \bar{R}$, be normally distributed with μ_3 and standard deviation σ_3. Similar remarks apply to the next three distributions. However, for ease of presentation, we choose the approach as presented here.

	Variable	Population	Mean	Standard Deviation
For each Q,	$f(Q,D)$	$D \in \bar{R}$	μ_1	σ_1 (see figure 3(a))
For each Q,	$f(Q,D')$	$D' \in \bar{I}$	μ_2	σ_2 (see figure 3(a))
For each $D_1 \in R$,	$f(D_1,D_2)$	$D_2 \in \bar{R}$	μ_3	σ_3
For each $D_1' \in I$,	$f(D_1',D_2)$	$D_2 \in \bar{R}$	μ_4	σ_4
For each $D_1' \in I$,	$f(D_1',D_2')$	$D_2' \in \bar{I}$	μ_5	σ_5
For each $D_1 \in R$,	$f(D_1,D_2')$	$D_2' \in \bar{I}$	μ_6	σ_6

The last four density functions can be obtained from the first two by proper substitutions.

As in the Swets model [Swets 1963, Brookes 1968], we have made two rather strict assumptions in the above discussion, namely, the distributions are continuous and are normal. The distributions may be approximated by continuous curves if the collection size is very large. By the Central Limit Theorem [Feller 1967], it may be argued that the distributions are normal. These assumptions are recently questioned by some authors [Heine 1974, Bookstein 1974]. Specifically, Heine [1974] doubts that the distributions are normal. No extensive experiments have been performed to validate or falsify the assumption. However, Heine admits that "simulation studies carried out indicate that the assumption is not seriously in error." Moreover, the experimental results by Swets [1969] and the explanation by Brookes indicate that "the probability density functions

are Gaussian (normal) or very nearly so.".

Based on the above definitions, the total numbers of relevant and irrelevant documents retrieved by Q at threshold value T are

$$\frac{k_1}{\sqrt{2\pi}\,\sigma_1}\int_T^\infty \exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right)dx \quad \text{and} \quad \frac{k_2}{\sqrt{2\pi}\,\sigma_2}\int_T^\infty \exp\left(-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right)dx$$

respectively, where k_1 is the number of documents in $\bar{R}$ and k_2 is the number of documents in $\bar{I}$.

If the set of index terms representing the document are chosen appropriately, the set of relevant documents should be close together while the irrelevant documents are dispersed in the space. Mathematically, this means that $\mu_3 > \mu_6$. Let D_2 be a relevant document relative to Q and D_1' be an irrelevant document retrieved by Q. D_2 may or may not be retrieved by Q. In the former case, Q and D_2 have quite a few terms in common. Similarly, there are many terms in common between D_1' and Q. The fact that D_1' and D_2 are classified in different categories with respect to Q makes it extremely unlikely that D_1' has a high correlation with D_2, assuming that the relevant assessment is correct. (If two documents are close to a given query, it is not necessarily true that they are close to one another). If D_2 is not retrieved by Q, then there is practically no relation between D_1' and D_2. Hence, in either case, the correlation of D_1' with a "random" document $D_2 \varepsilon \bar{R}$ is about the same as that of D_1' with an arbitrary random document.

Thus, the average value of $f(D_1', D_2), D_2 \varepsilon \bar{R}$, could be the same as that of $f(D_1', D_2')$, $D_2' \varepsilon \bar{I} - \{D_1'\}$. Since D_1' belongs to $\bar{I}$, the situation $D_2' = D_1'$ in the distribution of $f(D_1', D_2')$, $D_2' \varepsilon \bar{I}$ (please refer to the fifth distribution of the Table) must occur. Thus, the average value of $f(D_1', D_2')$, $D_2' \varepsilon \bar{I}$ is slightly greater than that of $f(D_1', D_2')$, $D_2' \varepsilon \bar{I} - \{D'\}$. On the other hand, the situation $D_2 = D_1'$ in the distribution of $f(D_1', D_2)$, $D_2 \varepsilon \bar{R}$ (please refer to the fourth distribution of the Table) can never arise. As a consequence, the average value of $f(D_1', D_2')$ $D_2' \varepsilon \bar{I}$ is greater than that of $f(D_1', D_2)$, $D_2 \varepsilon \bar{R}$ (i.e. $\mu_5 > \mu_4$), though it is expected that the difference is really very small. If the query Q is properly formulated, we may expect that it is on the average closer to the relevant documents than to the irrelevant ones. It follows that $\mu_1 > \mu_2$. Hence, throughout this paper, it is assumed that $\mu_1 > \mu_2, \mu_3 > \mu_6$ and $\mu_5 > \mu_4$.

Taking the inner product of a relevant document $D \varepsilon \bar{R}$ with both sides of (2.1), we get

$$f(Q^*, D) = f(Q, D) + \alpha \sum_{D_i \varepsilon R} f(D_i, D) - \beta \sum_{D_i' \varepsilon I} f(D_i', D). \tag{2.2}$$

Hence, assuming the mutual independence of the variables on the right side of (2.2), the expected value μ_1^* and standard deviation μ_1^* of $f(Q^*, D)$, $D \varepsilon \bar{R}$, can be shown to be

$$\mu_1^* \equiv \mu_1 + \alpha r \mu_3 - \beta s \mu_4$$

$$\text{and} \quad \sigma_1^* \equiv (\sigma_1^2 + \alpha^2 r \sigma_3^2 + \beta^2 s \sigma_4^2)^{1/2} \tag{2.3}$$

respectively, where r and s are the numbers of documents of R and I respectively. According to probability theory [Feller 1967], $f(Q^*,D)$ is also normally distributed. Similarly, the variable $f(Q^*,D')$, $D' \varepsilon \bar{I}$, is normally distributed with expected value μ_2^* and standard deviation σ_2^*, where

$$\mu_2^* \equiv \mu_2 + \alpha r \mu_6 - \beta s \mu_5$$

$$\text{and} \quad \sigma_2^* \equiv (\sigma^2 + \alpha^2 r \sigma_6^2 + \beta^2 s \sigma_5^2)^{1/2} \qquad (2.4)$$

The above relations can be shown by the Figures 3(a) and 3(b). Figure 3(a) shows the distribution of the relevant and irrelevant documents with respect to the original query Q. It is seen that a relevant document is closer to Q than an irrelevant document by an average "distance" of $(\mu_1-\mu_2)$. Clearly, if the dispersions (standard deviation) of the relevant and irrelevant documents remain unchanged while the separation of the relevant documents from the irrelevant document increases, better retrieval result is expected. Figure 3(b) shows the distribution of the documents with respect to the new query Q*. The new "distance" between the relevant documents and irrelevant documents has been increased to $\mu_1^*-\mu_2^* \equiv (\mu_1-\mu_2) + \alpha r(\mu_3-\mu_6) + \beta s(\mu_5-\mu_4)$. Unfortunately, the dispersions of the documents may have increased too. It is not clear from these two figures that Q* performs better than Q.

Thus, we have introduced a probabilistic model for the

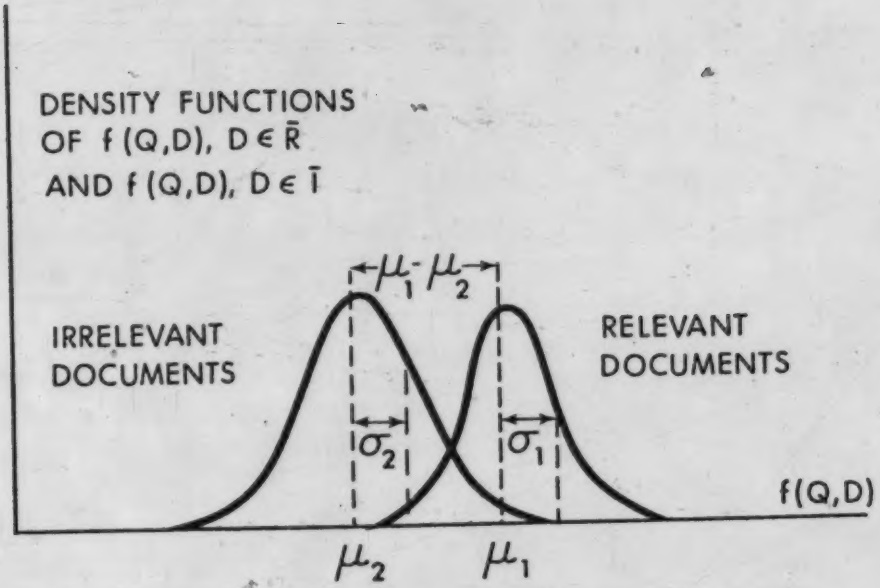

Fig. 3(a) The distributions of the relevant and irrelevant documents with respect to the original query.

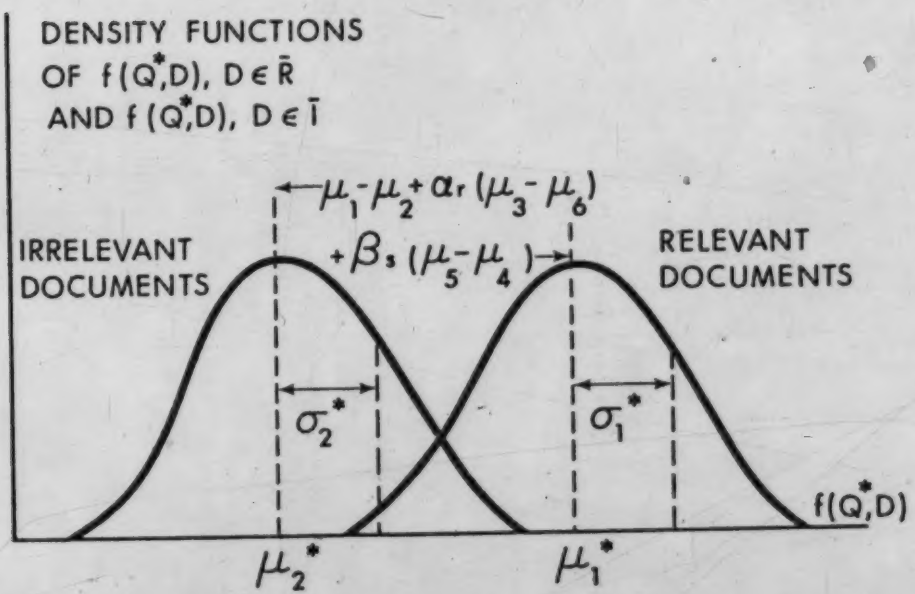

Fig. 3(b) The distributions of the relevant and the irrelevant documents with respect to the modified query.

distributions of the documents with respect to a query. In the next section, based on this model,we shall compare the retrieval performances of Q and Q*. In particular, some necessary and sufficient conditions on α and β will be derived such that Q* is better than Q. A region in the (α,β)-plane is found. Any point of this region lengthens the "distance" between the relevant documents and the irrelevant documents and makes sure that the irrelevant (Corollary 2.2, Condition 4) documents are no more important than the relevant documents in formulating the modified query.

2.3 Some Necessary and Sufficient Conditions

Let T* be the threshold when Q* is used to retrieve documents. For a fair comparison between Q and Q* in retrieval performance, the same number of documents should be retrieved by both of them. Thus the following relation holds:

$$\frac{k_1}{\sigma_1^*}\int_{T^*(\alpha,\beta)}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_1^*}{\sigma_1^*}\right)^2\right)dx+\frac{k_2}{\sigma_2^*}\int_{T^*(\alpha,\beta)}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_2^*}{\sigma_2^*}\right)^2\right)dx$$

$$= \frac{k_1}{\sigma_1}\int_{T}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right)dx + \frac{k_2}{\sigma_2}\int_{T}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right)dx \qquad (2.5)$$

As α or β changes, T* varies so that (2.5) holds. Hence, (2.5) may be regarded as an equation which defines

T^* as a continuous function of α and β. Moreover, we have

$$\lim_{\substack{\alpha\to 0\\ \beta\to 0}} T^*(\alpha,\beta) = T \tag{2.6}$$

Definition. The retrieval performance of Q^* is said to be as good as Q, or notationally $Q^* \geq Q$, if and only if Q^* retrieves at least as many relevant documents as Q while the total number of documents retrieved by both queries is the same. If Q^* retrieves more relevant documents than Q while the same number of documents are retrieved, then we write $Q^* > Q$.

The following theorem states a necessary and sufficient condition for $Q^* \geq Q$.

Theorem 2.1 Let Q^* be obtained from Q by (2.1), then $Q^* \geq Q$ if and only if (α,β) satisfy the following inequality

$$(T-\mu_1)\sigma_1^*/\sigma_1 + \mu_1^* \geq (T-\mu_2)\sigma_2^*/\sigma_2 + \mu_2^* \tag{2.7}$$

where μ_1^*, σ_1^*, μ_2^* and σ_2^* are defined as in (2.3) and (2.4).

Proof: By applying the transformations $y = (x-\mu_i^*)/\sigma_i^*$, $i = 1,2$, and $y = (x-\mu_i)/\sigma_i$, $i = 1,2$, to the left and right side of (2.5) respectively, we get

$$k_1 \int_{g_1}^{\infty} \exp(-y^2/2)dy + k_2 \int_{g_2}^{\infty} \exp(-y^2/2)dy$$

$$= k_1 \int_{(T-\mu_1)/\sigma_1}^{\infty} \exp(-y^2/2)dy + k_2 \int_{(T-\mu_2)/\sigma_2}^{\infty} \exp(-y^2/2)dy,$$

where $g_i = (T^*(\alpha,\beta)-\mu_i^*)/\sigma_i^*$, $i = 1,2$. (2.8)

We first prove the necessary condition. Suppose $Q^* \geq Q$, then $g_1 \leq (T-\mu_1)/\sigma_1$. By (2.8), we have $g_2 \geq (T-\mu_2)/\sigma_2$. Hence $T^* \leq (T-\mu_1)\,\sigma_1^*/\sigma_1 + \mu_1^*$ and $T^* \geq (T-\mu_2)\sigma_2^*/\sigma_2 + \mu_2^*$. The last two inequalities imply (2.7). Similarly, the sufficient condition can be proved by contradiction.

Although (2.7) is a necessary and sufficient condition for $Q^* \geq Q$, it is difficult to visualize a region in the (α,β)-plane which satisfies (2.7). In order to facilitate further discussion on this aspect, let us first parametrize α and β by two new variables m and t such that

$$\alpha r = mt \text{ and } \beta s = t, \quad m \geq 0, \; t \geq 0 \tag{2.9}$$

With the new parameters, (2.1), (2.3) and (2.4) can be reexpressed as

$$Q^* = Q + mt \sum_{D_i \in R} D_i/r - t \sum_{D_i' \in I} D_i'/s, \tag{2.10}$$

$$\mu_1^* = \mu_1 + mt\mu_3 - t\mu_4, \quad \sigma_1^* = (\sigma_1^2 + m^2t^2\sigma_3^2/r + t^2\sigma_4^2/s)^{1/2}, \text{ and}$$

$$\mu_2^* = \mu_2 + mt\mu_6 - t\mu_5, \quad \sigma_2^* = (\sigma_2^2 + m^2t^2\sigma_6^2/r + t^2\sigma_5^2/s)^{1/2} \tag{2.11}$$

We now elaborate on the geometry of the (α,β)-plane, as applied in this paper. The plane is meant to be the positive quadrant i.e. $\alpha \geq 0$, $\beta \geq 0$. A point in the plane is either represented by the rectangular co-ordinate (α,β) or by the parameter (m,t). To avoid the difficulties in representing points near the α-axis in the latter form, (2.9) can be rewritten as

$$\alpha r = t, \quad s = t/m, \quad m>1, \quad t\geq 0 \qquad (2.9a)$$

There are two forms of lines used in this paper: one is characterized by a fixed α (or β) and the other by a fixed m. The former is a line parallel to the β-axis (or the α-axis) and the latter is a line through the origin with slope $r/(ms)$.[†] For our future discussion, a region in the plane is either a rectangle bounded by four lines parallel to the α and the β-axis or is a sector bounded by two lines through the origin.

According to (2.1), a modified query Q* is uniquely determined by the values of α and β, given the original query Q. Hence, there is a one - one correspondence between the (α,β)-plane and the set of all modified queries defined by (2.1). Any geometrical entities (e.g. points, lines, regions) are a subset of the (α,β) plane and hence can be interpreted as a subset of the modified queries.

The following corollary gives the minimal possible value of m such that the whole half-line defined by (2.9) satisfies (2.7).

Corollary 2.2. Let the number of relevant documents and the number of irrelevant documents retrieved by Q be non-zero, i.e. $r\neq 0$ and $s\neq 0$. Let Q* be obtained by (2.10) and have standard deviations σ_1^* and σ_2^* as given by (2.11). If

---

[†]The α-axis and the β-axis are represented by $m = \infty$ and $m = 0$ respectively.

the following hypotheses are true, then $Q^* \geq Q$.

(1) The number of documents retrieved by Q at the threshold T is no more than half of the total number of documents.

(2) $\sigma_1^*/\sigma_1 \geq \sigma_2^*/\sigma_2$.

(3) $\sigma_1(\mu_3-\mu_6) > \sigma_3(\mu_1-\mu_2)/\sqrt{r}$

(4) $$m^2 \geq \frac{\sigma_4^2(\mu_1-\mu_2)^2/s-(\mu_5-\mu_4)^2\sigma_1^2}{(\mu_3-\mu_6)^2\sigma_1^2-\sigma_3^2(\mu_1-\mu_2)^2/r} .$$

Proof. By (2.11) and the inequality $[1+x^2]^{1/2} - 1 \leq x$ for $x \geq 0$, we have (by letting $x^2 = m^2t^2\sigma_3^2/(r\sigma_1^2) + t^2\sigma_4^2/(s\sigma_1^2)$)

$$[(T-\mu_1)\sigma_1^*/\sigma_1 + \mu_1^*] - [(T-\mu_2)\sigma_2^*/\sigma_2 + \mu_2^*]$$

$$= (\mu_1^*-\mu_2^*) - (\mu_1-\mu_2)\sigma_1^*/\sigma_1 + \{(T-\mu_2)(\sigma_1^*/\sigma_1-\sigma_2^*/\sigma_2)\}$$

$$= mt(\mu_3-\mu_6)+t(\mu_5-\mu_4)-(\mu_1-\mu_2)\{[1+m^2t^2\sigma_3^2/(r\sigma_1^2)+t^2\sigma_4^2/(s\sigma_1^2)]^{1/2}-1\}$$

$$+ \{(T-\mu_2)(\sigma_1^*/\sigma_1 - \sigma_2^*/\sigma_2)\}$$

$$\geq [m(\mu_3-\mu_6) + (\mu_5-\mu_4) - (\mu_1-\mu_2)(m^2\sigma_3^2/(r\sigma_1^2)+\sigma_4^2/(s\sigma_1^2))^{1/2}]t$$

$$+ [(T-\mu_2)(\sigma_1^*/\sigma_1 - \sigma_2^*/\sigma_2)] \qquad (2.12)$$

By Hypotheses (3) and (4), we get $\sigma_1^2[m(\mu_3-\mu_6)+(\mu_5-\mu_4)]^2 \geq \alpha_1^2[m^2(\mu_3-\mu_6)^2 + (\mu_5-\mu_4)^2] \geq (\mu_1-\mu_2)^2(m^2\sigma_3^2/r + \sigma_4^2/s)$. Taking square root on both sides, it shows that the first bracketed expression of (2.12) is non-negative.

Next, we claim that $T \geq \mu_2$. Otherwise we would have $T < \mu_2 \leq \mu_1$. Then, the lower limits of the two integrals on the right-hand side of (2.8) have negative values, implying that Q retrieves more than half of the total number of documents, contrary to Hypothesis (2), the second bracketed expression of (2.12) is non-negative.

It follows from Theorem 2.1 that $Q^* \geq Q$.

Remarks. This theorem suggests, within the framework of this model, a rather practical way to improve retrieval results,provided that the four hypotheses are true. Hypothesis (4) specifies the sector in the (α,β) plane whose points define better queries in retrieval performance. (The sector is bounded by two lines. One is the α-axis. The other is a line through the origin whose "slope" m satisfies hypothesis (4)). Now, we shall attempt to examine to what extent hypotheses (1) - (3) are realistic.

Usually, a query retrieves only a small portion of the documents and thus hypothesis (1) probably holds under normal retrieval environment. Owing to the lack of information concerning the standard deviations,it is harder to justify hypotheses (2) and (3). If the collection of documents are properly indexed, it is expected that the average closeness of two relevant documents relative to a given query is a lot larger than that of a relevant document with an irrelevant document. Thus, $\mu_3 >> \mu_6$. For a query which requires the feedback operation, we may argue that μ_1 can-

not be much larger than μ_2. Since a document usually has more terms than a query, the assumption that μ_3 is not less than μ_1 is justified. As a consequence, $\mu_3-\mu_6 > \mu_1-\mu_2$. Thus, a case in which hypotheses (2) and (3) hold is that $\sigma_i = \sigma_j$, $1 \le i,j \le 6$. It is obvious that the above process can be iterated to produce better and better queries i.e. if Q^* is obtained from Q by (α_1,β_1), then Q^{**} can be gotten from Q^* by (α_2,β_2) with $Q^{**} \ge Q^* \ge Q$ and so on. Since $\mu_1^* \neq \mu_1$ and $\mu_2^* \neq \mu_2$, it may be necessary to choose (α_2,β_2) such that $\alpha_2 \neq \alpha_1$ or $\beta_2 \neq \beta_1$.

Sometimes the original query Q is formulated so badly by the user that not even a single relevant document is retrieved. Under such a condition, no value of α would help improve the retrieval performance. The following corollary states that if a small value for β is specified in (2.1), then $Q^* > Q$.

Corollary 2.3. Let the number of relevant documents retrieved by Q be zero, i.e. $r = 0$, and Q^* be obtained by (2.1). If the following hypotheses are true, then $Q^* > Q$.

(1) The number of documents retrieved by Q at the threshold T is no more than half of the total number of documents.

(2) $\sigma_4\sigma_2 \ge \sigma_1\sigma_5$ and $\sigma_4(\mu_1-\mu_2) > \sqrt{s}\,\sigma_1\,(\mu_5-\mu_4)$

(3) α is arbitrary and

$$\beta \le 2(\mu_5-\mu_4)(\mu_1-\mu_2)\sigma_1^2/[\sigma_4^2(\mu_1-\mu_2)^2/s - \sigma_1^2(\mu_5-\mu_4)^2]$$

Proof: Similar to that of Corollary 2.2.

The last few corollaries are based on a number of assumptions about the expected values and standard deviations of the normal distributions. Our next result does not depend on these assumptions.

Theorem 2.4. Let the number of relevant documents and the number of irrelevant documents retrieved by Q be non-zero. Let Q* be obtained from Q by (2.1). If α and β are sufficiently small, then Q*>Q.

Proof. Let any point (α,β) in the positive quadrant of the (α,β)-plane be parametrized by (2.9). We first consider the case $1 \geq m \geq 0$. Substitution of (2.9) into (2.7) gives

$$k_1 \int_{g_1}^{\infty} \exp(-y^2/2)dy + k_2 \int_{g_2}^{\infty} \exp(-y^2/2)dy \tag{2.13}$$

$$= k_1 \int_{(T-\mu_1)/\sigma_1}^{\infty} \exp(-y^2/2)dy + k_2 \int_{(T-\mu_2)/\sigma_2}^{\infty} \exp(-y^2/2)dy$$

where $g_i(m,t) = (T^*(m,t)-\mu_i^*)/\sigma_i^*$, $i = 1,2$, and μ_1^*, μ_2^*, σ_1^* σ_2^* are defined as in (2.11).

The number of relevant documents retrieved by Q* at the threshold T*(m,t) is

$$I(m,t) \equiv (k_1/\sqrt{2\pi}) \int_{g_1(m,t)}^{\infty} \exp(-y^2/2)dy. \tag{2.14}$$

The expression of $\partial T^*(m,t)/\partial t$ can be obtained by differentiating equation (2.13) with respect to t. Substituting this expression to $\partial I/\partial t$, we get

$$\partial I/\partial t = \{(k_1k_2G_1G_2)/[\sqrt{2\pi}\ (\sigma_1^*)^3(\sigma_2^*)^3(G_1/\sigma_1^*+G_2/\sigma_2^*)]\}$$

$$\{-t^3(\mu_1-\mu_2)(m^2\sigma_6^2/r + \sigma_5^2/s)(m^2\sigma_3^2/r + \sigma_4^2/s)$$

$$+ t^2[\sigma_2^2(\mu_5-m\mu_6)(m^2\sigma_3^2/r + \sigma_4^2/s) + \sigma_1^2(m\mu_3-\mu_4)(m^2\sigma_6^2/r + \sigma_5^2/s)]$$

$$- t[(T^*(m,t)-\mu_2)(m^2\sigma_6^2/r + \sigma_5^2/s)\sigma_1^2 - (T^*(m,t)-\mu_1)(m^2\sigma_3^2/r + \sigma_4^2/s)\sigma_2^2]$$

$$+ \sigma_1^2\sigma_2^2[(\mu_5-\mu_4) + m(\mu_3-\mu_6)]\} \qquad (2.15)$$

where $G_i = \exp(-\frac{1}{2}(\frac{T^*(m,t)-\mu_i^*}{\sigma_i^*})^2)$, $i = 1,2$.

By (2.6), $T^*(m,t)\to T$ as $t\to 0$. It is also obvious that, for a fixed value of m, G_i and σ_i^*, $i=1,2$, are positive and are bounded as t tends to 0. Thus, $\partial I(m,t)/\partial t > 0$ for sufficiently small value of t. It then follows that $Q^*>Q$ for sufficiently small non-negative values of α and β (with at least one of them strictly positive) in the sector defined by $\beta\geq 0$ and $\alpha r \leq \beta s$.

For the case where $m>1$, we can set $n = 1/m$ and parametrize (α,β) by $\alpha r = t$, $\beta s = nt$ instead of (2.9). Proceeding simlarly as above, we can then show that $Q^*>Q$ for sufficiently small non-negative values of α and β in the sector defined by $\beta\geq 0$ and $\alpha r \geq \beta s$.

2.4 Optimal Values for α and β

We now attempt to find the point (α,β) in the (α,β)-plane which maximizes the performance of the new query Q^*. Although we do not succeed in getting a closed form formula

for the point, it is found that the point must lie on a hyperbolic curve. We shall first show that the point must lie in a finite region of the plane.

Using the notation of the last section, it will be shown that for sufficiently large t, I(m,t) is decreasing for any value of m such that $\infty > m \geq 0$. Furthermore, the optimal point cannot lie on the α axis and the β axis. Three technical lemmas 2.5-2.7 lead to the results stated in Theorem 2.8. For their proofs, please refer to Appendix I.

Lemma 2.5. There exists a constant $c_1>0$ and a constant $t_1>0$ such that for every $t \geq t_1$, every $0 \leq m \leq 1$, the threshold $T^*(m,t) \leq c_1 t$.

Lemma 2.6. There exists a constant t_5 such that for every $t \geq t_5$ and for $0 \leq m$, $\partial I/\partial t(m,t) < 0$.

Instead of expressing I in terms of m and t, we may write I as a function of α and β by means of (2.9) and (2.14), i.e.,

$$I(\alpha,\beta) = (k_1/\sqrt{2\pi}) \int_{g_1(\alpha,\beta)}^{\infty} \exp(-y^2/2)dy$$

Lemma 2.7. $\partial I/\partial\alpha(0,\beta)>0$ and $\partial I/\partial\beta(\alpha,0)>0$ when $r,s \neq 0$.

Theorem 2.8. Let the number of relevant documents and the number of irrelevant documents retrieved by the original query Q be r and s respectively, with $r,s \neq 0$. Let $(\alpha_1,\beta_1) = (t_5/r, t_5/s)$, where t_5 is determined in Lemma 2.6. Then there exists a point (α_2,β_2) with $0<\alpha_2<\alpha_1$ and $0<\beta_2<\beta_1$ such that the Q^* defined by the parameter (α_2,β_2) is at least as good as the Q^* defined any (α,β), $\alpha,\beta \geq 0$. Furthermore, the point

(α_2,β_2) must lie on the curve

$$\alpha(\sigma_6^2\sigma_1^2-\sigma_3^2\sigma_2^2)(\mu_5-\mu_4) + \beta(\sigma_4^2\sigma_2^2-\sigma_5^2\sigma_1^2)(\mu_3-\mu_6)$$
$$= (\alpha\beta)(\sigma_6^2\sigma_4^2-\sigma_5^2\sigma_3^2)(\mu_1-\mu_2) \qquad (2.16)$$

Proof. Let W be the closed region bounded by the lines

$$L_1: \ \alpha=0, \ 0\le\beta\le\beta_1,$$
$$L_2: \ \beta=0, \ 0\le\alpha\le\alpha_1,$$
$$L_3: \ \alpha=\alpha_1, \ 0\le\beta\le\beta_1,$$
and $$L_4: \ \beta=\beta_1, \ 0\le\alpha\le\alpha_1.$$

Let the maximum of $I(\alpha,\beta)$ in W occur at (α_2,β_2). We now show that $I(\alpha_2,\beta_2) \ge I(\alpha,\beta)$ for any $\alpha\ge0$, $\beta\ge0$ and (α_2,β_2) lies in the interior of W.

Let (α_5,β_5) be any point in the quadrant $\alpha\ge0$, $\beta\ge0$ but be outside of W. Then $\alpha_5>\alpha_1$ or $\beta_5>\beta_1$. Let $m_5 = (\alpha_5 r)/(\beta_5 s)$ and L_5 be the line connecting (α_5,β_5) and the origin. L_5 intersects either L_4 or L_3. Without loss of generality, suppose L_5 intersects L_4 at (α_4,β_1). L_5 can be parametrized by $\alpha = m_5 t/r$ and $\beta = t/s$ where $0\le t\le\beta_5 s$. In particular, the portion of the line from (α_4,β_1) to (α_5,β_5) is defined by $t_5\le t\le\beta_5 s$. By Lemma 2.6, $I(m_5,t)$ is decreasing along this portion of L_5. Thus, $I(m_5,\beta_5 s) < I(m_5,t_5)$ i.e., $I(\alpha_5,\beta_5) < I(\alpha_4,\beta_1)$. Since $I(\alpha_4,\beta_1) \le I(\alpha_2,\beta_2)$, it follows that $I(\alpha_5,\beta_5) < I(\alpha_2,\beta_2)$.

By Lemma 2.7, it is obvious that $I(\alpha,\beta)$ cannot attain its maximum of L_1 or L_2. To show that (α_2,β_2) cannot lie on L_4, consider any point (m_4,t_5) on L_4 where $m_4>0$. Since $\partial I(m,t)/\partial t$ is negative and is continuous at (m_4,t_5),

there must be a neighborhood of (m_4,t_5) within which $\partial I(m,t)/\partial t$ is negative. Thus there exists $t_6<t_5$ such that, $\partial I(m,t)/\partial t$ is negative on $t_6<t<t_5$, $m = m_4$. Hence $I(m_4,t_5) < I(m_4,t)$, $t_6<t<t_5$. But $m = m_4$, $t_6<t<t_5$ are interior points of W. Thus, (α_2,β_2) cannot lie on L_4. Similarly, (α_2,β_2) cannot lie on L_3. As a consequence, (α_2,β_2) satisfies $\partial I/\partial\alpha(\alpha,\beta) = \partial I/\partial\beta(\alpha,\beta) = 0$. From $\partial I/\partial\alpha(\alpha,\beta) = 0$, we obtain

$$T^*(\alpha,\beta) = \{1/\alpha(\sigma_6^2(\sigma_1^*)^2-\sigma_3^2(\sigma_2^*)^2)\}\cdot\{(\mu_3-\mu_6)(\sigma_1^*)^2(\sigma_2^*)^2 - \alpha(\mu_1^*\sigma_3^2(\sigma_2^*)^2 - \mu_2^*\sigma_6^2(\sigma_1^*)^2)\}. \quad (2.17)$$

Similarly from $\partial I/\partial\beta(\alpha,\beta) = 0$, we find

$$T^*(\alpha,\beta) = \{1/\beta(\sigma_5^2(\sigma_1^*)^2-\sigma_4^2(\sigma_2^*)^2)\}\cdot\{(\mu_5-\mu_4)(\sigma_1^*)^2(\sigma_2^*)^2 - \beta(\mu_1^*\sigma_4^2(\sigma_2^*)^2 - \mu_2^*\sigma_5^2(\sigma_1^*)^2)\} \quad (2.18)$$

Equating (2.17) and (2.18), (2.16) follows.

A numerical solution for (α_2,β_2) can be obtained by substituting the relation between α_2 and β_2 and equation (2.16) into equation (2.13).

2.5 Experimental Results

Experiments are performed on a collection of 200 documents on aerodynamics (CRN2NULDOCS 200). The 42 queries (CRN2NUL Quests 42) are used to retrieve the documents by means of the simple matching function defined in Section 1.1. For each query, the threshold is set such that the first ten documents which correlate highest with the query are retrieved. Of the 42 queries, it is found that 8 queries do not retrieve any relevant document. The first 12 queries, each retrieving at least one relevant document

are chosen. A number of (α,β) values are selected according to corollary 2.2 i.e. the new queries defined by the (α,β)'s should be at least as good as the original queries. With the absence of information about the standard deviations and the expected values of the random variables defined in section 1.1, we arbitrarily divide the α-β plane into two parts by the line $\alpha r = \beta s$. The set of (α,β) values tested satisfy $\alpha r \geq \beta s$. The exact (α,β) values are shown in figure 4(a). Each point in the diagram is represented by a number from 1 to 5. The average performance of the new queries by the different (α,β) values with respect to the original queries are shown in figure 4(b). The x-axis of the figure is the recall value averaged over the twelve queries, where recall is the proportion of relevant documents retrieved. The y-axis is the averaged precision value, where precision is the proportion of retrieved documents that are relevant. (detailed discussion of recall and precision can be found in [Salton 1968]). The '0' in the figure represents the performances of the original queries, while the performances of the new queries are indicated by numbers from 1 to 5 corresponding to the numbers assigned in figure 4(a). For example, the '1' represents the performance of the set of new queries defined by formula (2.1) with $\alpha = 100/r$ and $\beta = 1/s$. It is found that using any of these (α,β) values, all of the queries show some improvement over the original queries.

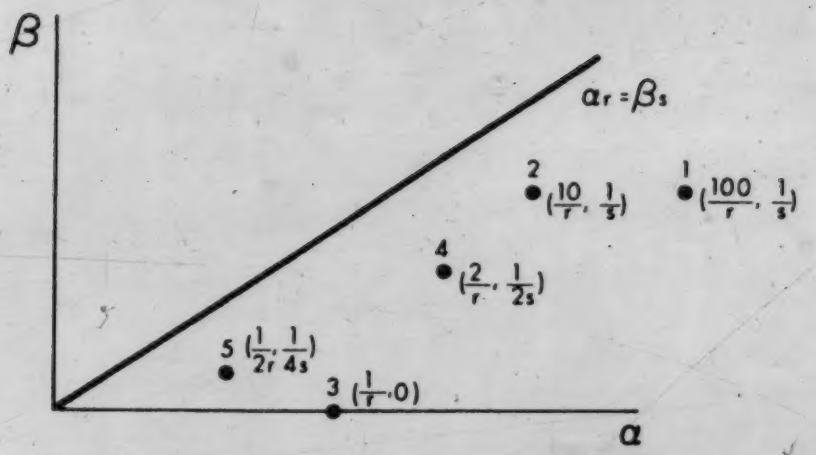

FIG. 4(a). Points on (α, β)-plane tested (not to scale)

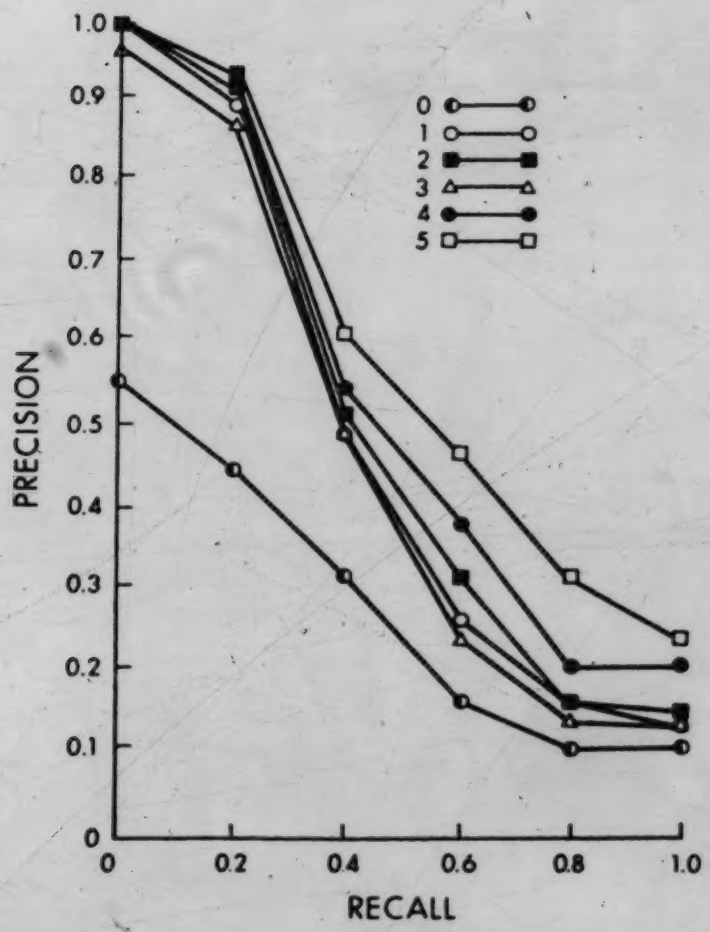

FIG. 4(b). Performance of the new queries (region of improvement)

To illustrate that not all (α,β)values with $\alpha\geq 0$, $\beta\geq 0$ define "good" new queries, a number of (α,β) values are selected in the other part of the plane i.e. $\alpha r>\beta s$. The exact (α,β) values are indicated in figure 5(a), with the performances of the corresponding queries in figure 5(b). It is found that of the five tested (α,β) values, only one defines better queries than the original ones. This particular value is closer to the line $\alpha r' = \beta s$ than the other points. While the line $\alpha r = \beta s$ may not be the exact line separating the parameters for defining "good" queries from those defining "bad" ones, it is seen that the experimental results obtained are consistent with the theoretical predictions of corollary 2.2.

In the case that the original queries do not retrieve any relevant document, it is sufficient to set the parameter β. The set of eight original queries which do not retrieve any relevant document are used to test which values of β should be chosen to form the new queries. According to corollary 2.3, β should not be set too high. A number of β values are indicated in figure 6(a) with the performances of the corresponding queries in figure 6(b). Since the original queries do not retrieve any relevant document, they have zero precision and are therefore represented by the x-axis. It is seen that with $\beta = 1/6s$ or $\beta = 1/5s$ i.e. small values of β, the improvement is highest.

It is suspected that the maximum improvement occurs at the intersection of the curve defined by equation (2.16)

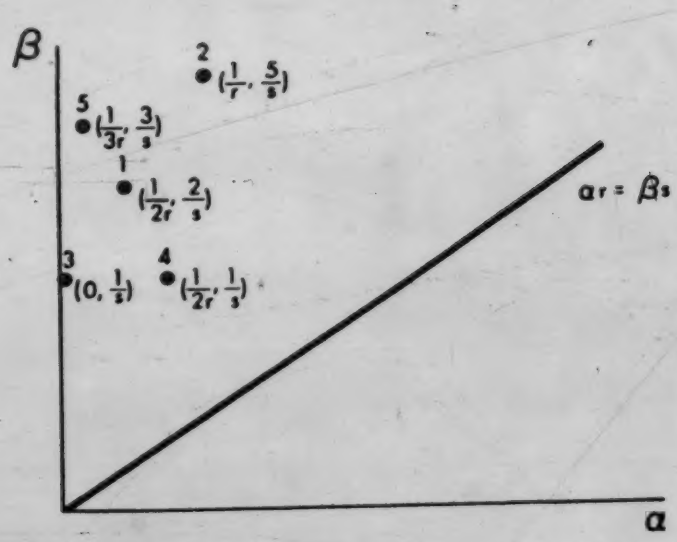

FIG. 5(a). Points on the $(\alpha - \beta)$-plane tested (not to scale)

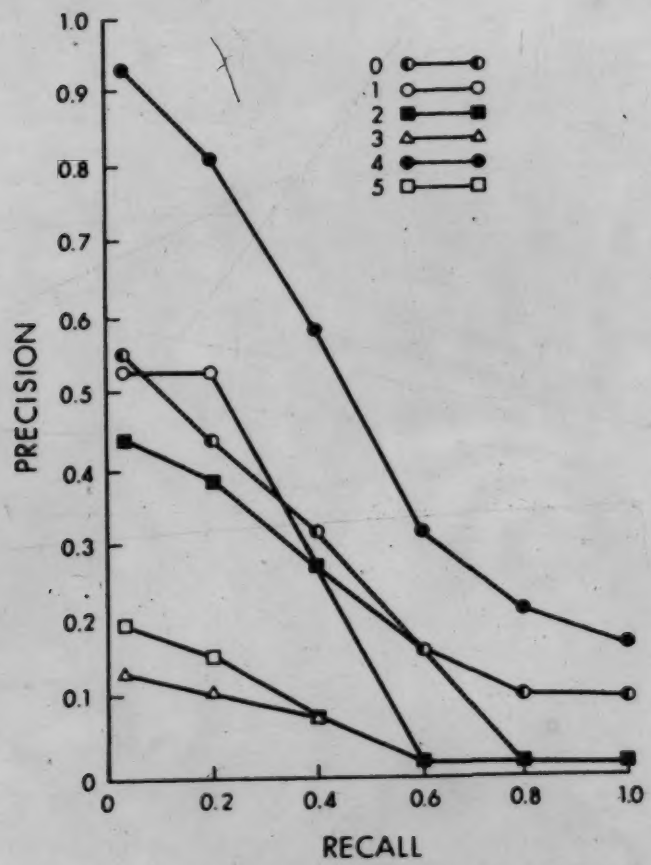

FIG. 5(b). Performance of the new queries (region of deterioration)

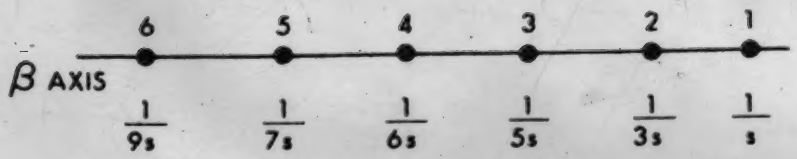

FIG. 6(a). Points on the β-axis tested (not in scale)

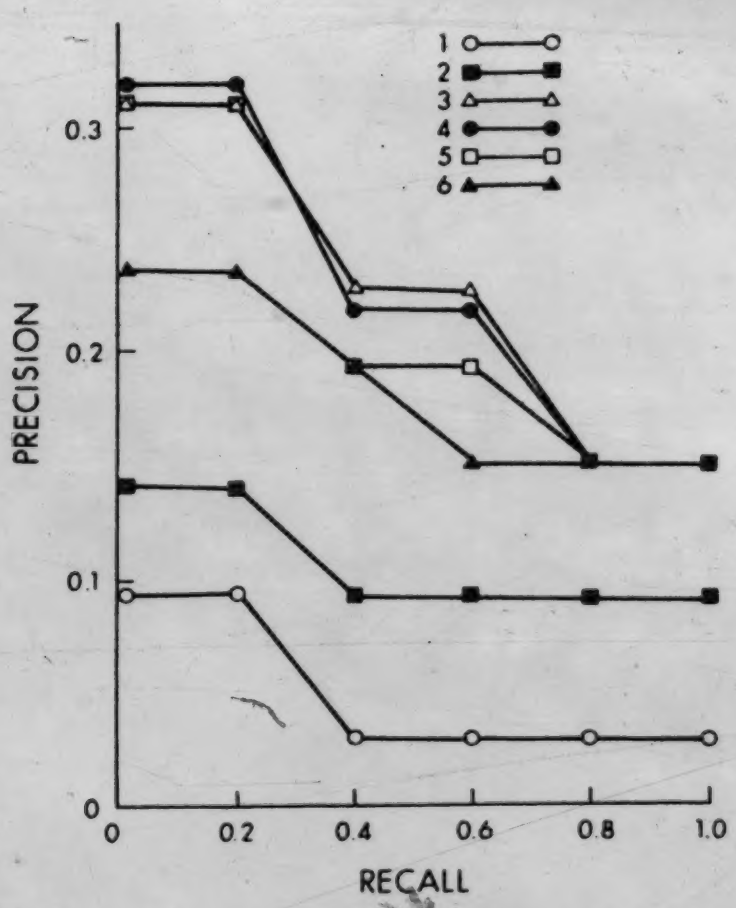

FIG. 6(b). Performance of the new queries (testing for the parameter β)

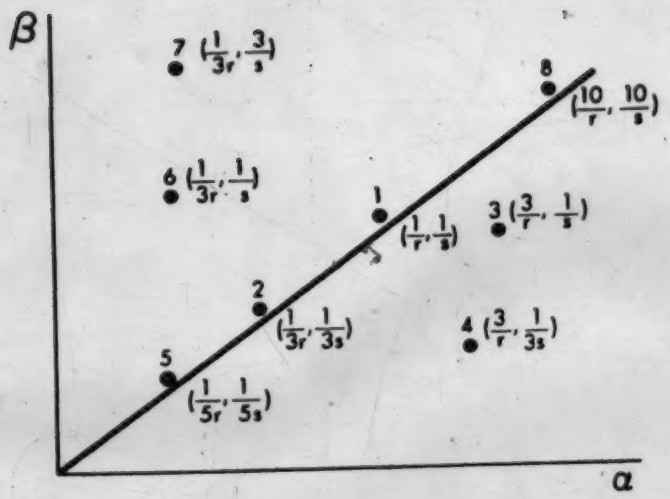

FIG. 7(a). Points on the (α, β)-plane tested (not to scale)

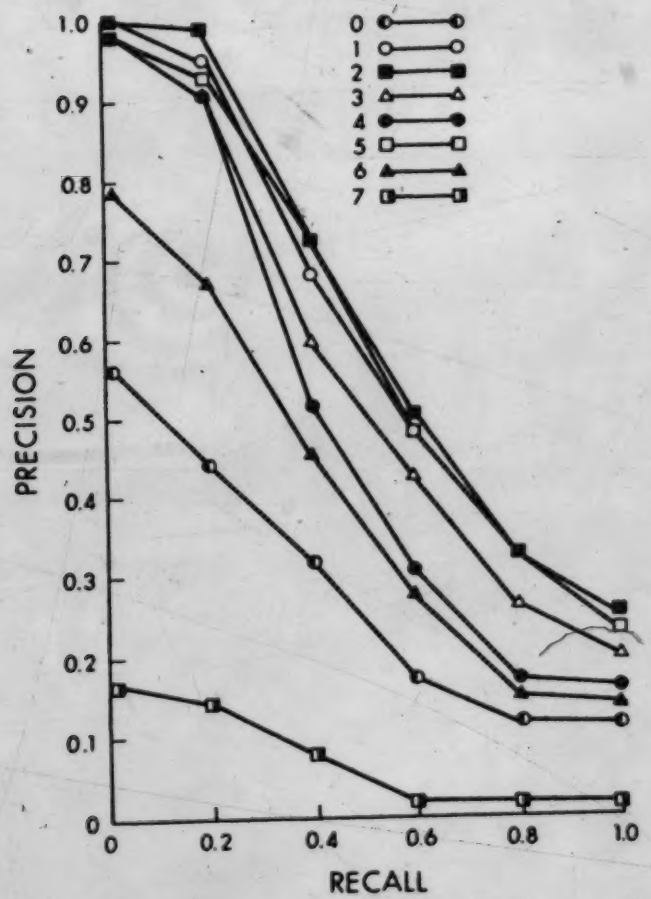

FIG. 7(b). Performance of the new queries (locating maximum)

and the line $\alpha r = \beta s$. A number of points on the line $\alpha r = \beta s$ as well as points in the neighborhood of (1/r,1/s) are tested. It is seen that the new queries with the parameters (1/3r,1/3s) exhibit the best overall performance, though the difference between them and the queries defined by the parameters (1/r,1/s) is very slight. The performances of the difference queries are illustrated in figure 7(b).

Although the validity of Cranfield relevance judgement has been questioned [Harter 1971, Swanson 1971], it is expected that as long as the relevant documents are "clustered" together and the irrelevant documents are "dispersed", similar retrieval results would be obtained for reasonable yet different relevance judgements. This is partially supported by our first set of experiments in which every new query modified by the parameters of figure 4(a) is at least as good as its corresponding original query. Furthermore, the new query as defined by (2.1) varies according to the relevance assessment.

2.6 Generalization

The analysis carried out in the previous sections is based on the simple matching function. It is easily seen that the approach can be generalized to any matching function ,

is the cosine function† used in the SMART system, equation (2.1) can be modified to become

$$Q^* = Q + \alpha \sum_{D_i \varepsilon R} D_i/|D_i| - \beta \sum_{D_i' \varepsilon I} D_i'/|D_i'|$$

where $|D_i|$ is the ℓ_2-norm of D_i. It then follows that $|Q^*|\cos(Q^*,D) = (Q^*,D/|D|) = (Q,D/|D|) + \alpha \sum_{D_i \varepsilon R} (D_i/|D_i|,D/|D|)$ $- \beta \sum_{D_i' \varepsilon I} (D_i'/|D_i'|,D/|D|)$. The approach used in the previous section can then be carried over in the case the cosine function is used, if we replace X by $X/|X|$ where X is a document.

---

†cosine $(X,Y) = (x,y)/(|x|\,|y|) = \dfrac{\sum x_i y_i}{|x||y|}$ where $x = (x_1,x_2,\ldots x_n)$ and $y = (y_1,y_2,\ldots y_n)$

CHAPTER 3

CLUSTERED FILES

3.1 Introduction

There are two types of search strategies (which have actually been implemented in the SMART system): Full search and cluster search. Full search is simple: the correlation (closeness or similarity) values between the query and all the records are calculated. For cluster search, the records have to be classified, by some clustering algorithms, into a number of sets, called clusters, with a representative constructed for each cluster. Instead of comparing the query vector with each record, Q is correlated with different representatives. Based on these correlation values, the system decides which clusters are to be searched. The data base can be arranged in a tree-like structure so that there might be sub-clusters within a cluster and so on. Now the calculation and subsequent ranking of records in order of the decreasing closeness are usually both laborious and time-consuming. As the trend now is for larger and more diversified data bases, cluster-based retrieval is distinctly more advantageous if the increase in efficiency does not incur serious loss of effectiveness. The loss is due to the fact that some of records should have been retrieved had full search been conducted on the whole data base, but are "missed" simply because they are stored in the part of data

base not searched.

The meaning of effectiveness is slightly different from what has been defined. No attempt is made to evaluate the relevance of the retrieved items as in chapter 2, and the concept of relevance does not apply here. Our intuition will have us believe that the higher the correlation value between an item and the query, the more probable that the item is relevant to the query. Indeed it is based on this principle that the best-match method was devised. However their exact relationship will not be elaborated here. We shall strictly adhere to the loss (as mentioned above) as the yardstick measuring the system effectiveness. It is our intention to do away with the human factor, i.e., relevance judgment and emphasize on the clustered files as used in a general context. For this reason, we will use "record" instead of "document" throughout this chapter. Corresponding to "relevant documents" with respect to a query Q, we define "desired records" as $\{O \mid f(Q,O) \geq k$, where k is some constant, O is a record$\}$.

3.2 Related Work

In reviewing the literatures on clustered-files, one can find most of them focusing on clustering or classification algorithms and their properties. Experiments [Burkhard et al 1973, Salton 1971] have been carried out on searching in cluster-based files. Generally they yield

reasonable retrieval performance. Rivest's thesis [Rivest 1973] examines a similar problem from the theoretical point of view and concludes that clustered search is most efficient among all "balanced hashing functions". However, the amount of computing time required is still very substantial. A further reduction in computing time can be achieved by examining only those clusters whose representatives are sufficiently close to the query Q. In many applications of on-line information retrieval, in particular in document retrieval or in situations where fast response time is required, it is sufficient to retrieve a majority of the desired records. It is therefore of great interest to obtain the percentage loss of desired records. With this information, the system manager (or the user who is "on-line") can decide on the trade-off between efficiency and effectiveness.

There is a wide variety of retrieval methods [van Rijsbergen 1974, Salton 1971]. Different classification algorithms specify clusters in different ways; for instance, some definitions of clusters permit a record to appear in more than one cluster, while other definitions forbid overlapping clusters. Clusters may either be organized in a tree-structure or they may occur in only one level. Here,a probabilistic model is constructed which contains all the essential characteristics of clusters as produced by a wide variety of different algorithms. It is

in this context that estimation of the number of desired records in one cluster with respect to that of another cluster is made, permitting an approximate percentage loss of desired records to be obtained. In this thesis, the variation of the ratio of the number of desired records in one cluster to that of another under changes of different parameters is considered. Empirical results are also obtained, based on which guidelines are provided for setting up the representatives of different clusters and searching the clusters.

The framework on which Rivest's analytical work is based differs substantially from that presented here. Specifically, the differences are as follows: (i) He uses a "distance" function which measures the number of attributes contained in one vector but not the other, instead of a "closeness" or "similarity" function. In some applications of information retrieval, the matching rather than the mismatching of the attributes is of importance; i.e., a distance function may not be an accurate inverse of similarity function in those applications. (ii) The search algorithm in Rivest's thesis obtains all the desired records at the expense of more retrieval time. Thus, there is no need to estimate the number of desired records in one set of clusters relative to that of another in his framework. (iii) It is assumed that the field is randomly chosen subset out of the 2^n possible records. This assumption is rather unrealistic in many applications of information retrieval.

Bentley [1975] attacks a similar problem using a multi-dimensional binary search tree. His method also obtains all the desired records at the expense of more computing time.

3.3 A Probabilistic Model

We now state the assumptions on which our analysis is based.

(i) The attributes in a cluster are independent; i.e., the occurrence of an attribute (or a set of attributes) has no relation with that of other attributes. Such an idealization is adopted by a number of authors in different contexts (e.g., [Bookstein 1975, Schkolnick 1975, Yu 1976]).

(ii) All records are assumed (conceptually) to be n-dimensional vectors, where n is the total number of attributes in the set of all records, and each attribute is either 0 or 1. However, the records may be stored by recording only the positions of the non-zero components.

With the above assumptions, the expected number of records in a cluster C of m records having i attributes in common with a query Q can be computed as follows. Let $\ell \geq i$ be the number of attributes (non-zero components) of Q. Let the attributes be denoted by $z_j, 1 \leq j \leq \ell$. The probability that any record of C contains z_j is $q_j = y_j/m$, where y_j is the number of records in the cluster C of m records having the jth attribute. Similarly, the probability that a record does not contain z_j is $(1-q_j)$. Using the assumptions, the

probability that a record contains $z_{j_1}, z_{j_2}, \ldots, z_{j_i}$ but not $z_{j_{i+1}}, \ldots, z_{j_\ell}$ is $\left(\prod_{k=1}^{i} g_{j_k}\right)\left(\prod_{k=i+1}^{\ell} (1-g_{j_k})\right)$. There are $\binom{\ell}{i}$ different ways of choosing i attributes out of ℓ. Thus, the probability that a record has exactly i attributes in common with Q is

$$\sum_{\binom{\ell}{i}} \left(\prod_{k=1}^{i} g_{j_k}\right)\left(\prod_{k=i+1}^{\ell} (1-g_{j_k})\right) \qquad (3.1)$$

where $\sum_{\binom{\ell}{i}}$ is summing over all the $\binom{\ell}{i}$ possible choices. (3.1) is obviously symmetric with respect to the y's. For the sake of simplicity, (3.1) can be denoted by $C(g_1, \ldots, g_\ell, i)$. The expected number of records having k or more attributes in common with Q is then given by $m \sum_{i=k}^{\ell} C(g_1, \ldots, g_\ell, i)$.

We now define the representative R of a given cluster $\mathcal{C}$.

Definition 3.1. Let $\mathcal{C}$ consist of the records $\{O_j\}$, $1 \le j \le m$; the jth record is given by the binary vector $O_j = (O_{j1}, O_{j2}, \ldots, O_{jn})$. Let $Y = (y_1, \ldots, y_n)$ be the (vector) sum of the m records, where $y_k = \sum_{j=1}^{m} O_{jk}$, $1 \le k \le n$. The kth component of the representative R is then defined to be 1 if $g_k = y_k/m \ge t$ (where $1 \le k \le n$ and t is any arbitrary number satisfying $0 \le t \le 1$) and 0 otherwise.

By the above definition, if an attribute occurs sufficiently after in the records of a cluster, then it will appear in the representative. There are two reasons for

defining the representative in this manner. Firstly, the amount of storage needed for such a representative is minimal. Those attributes which are unlikely to occur in a randomly selected record of the cluster (i.e. their probabilities of occurrences are less than t) are ignored. Only the positions of the non-zero components of the representative are actually stored. Secondly, the computation of the correlation of a query with a representative is efficient. The number of comparisons needed to find the number of attributes in common is at most equal to the sum of the numbers of the non-zero components of the two vectors, if the positions of the non-zero components of each vector are stored in ascending or descending order.

As mentioned in section 3.1, the user's query is compared with the different representatives and the correlation decides which clusters will be examined. Thus, it is necessary to relate the correlation of the query and the representative of a cluster with the number of records in the cluster having a fixed number of attributes in common with the query. Let the probabilities of occurrences of the ℓ attributes of Q in a randomly selected record of a cluster C be $(g_1, g_2, \ldots, g_\ell)$ and s be the number of attributes in common between Q and R, the representative of C. By the definition of a representative, s of the g's are greater than or equal to t and the other $(\ell-s)$ g's are less than t. Let us denote the probabilities by $(p_1, \ldots, p_s, q_{s+1}, \ldots, q_\ell)$

where each $p \geq t$ and each $q < t$. Thus, by our earlier discussion, the expected number of records in $\mathcal{C}$ having k or more attributes in common with Q is $m \sum_{i=k}^{\ell} C(p_1,\ldots,p_s,q_{s+1},\ldots,q_\ell,i)$. This expectation value is conditional on p_i and q_j, $1 \leq i \leq s$, $s+1 \leq j \leq \ell$. However, because of the way a representative is defined, the values of the p's and q's are unknown; at least this information cannot be readily obtained just by inspecting the representative. Thus, in the present context, it is more appropriate to consider the average behavior of clusters having the same characteristics (i.e. with their representatives having the same number of attributes in common with Q, the same number of records, and the same threshold t in the definition of the representative) than to examine the behavior of an individual cluster. By average is meant that the p's are allowed to vary independently from t to 1, and their distributions (which are not known) are identical. Similarly, the q's are assumed to be independent and identically distributed between 0 and t, though the distributions of a p and a q will be different. Thus, the "unconditional expected number" (as opposed to "conditional expected number" as introduced earlier) of records in $\mathcal{C}$ having k or more attributes in common with Q is $mE_{(p,q)} \{ \sum_{i=k}^{\ell} C(p_1,\ldots,p_s, q_{s+1},\ldots,q_\ell,i)$ where E is expected value over the independent random variables p's and q's, and m is the number of

records in $\mathcal{C}$.†

By probability theory [Feller 1967], $E(x_1+x_2)=E(x_1)+E(x_2)$ and, if x_1 and x_2 independent, $E(x_1x_2)=E(x_1)E(x_2)$. The above expression can then be reduced to $m \sum_{i=k}^{\ell} C(\bar{p}_1,\ldots,\bar{p}_s,\bar{q}_{s+1},\ldots,\bar{q}_\ell,i)$ where $\bar{p}_i=E(p)$ is the expected value of each p_i, $1\le i\le s$ and $\bar{q}_i=E(q)$ is the expected value of each q_i, $s+1\le i\le \ell$. In order to further simplify our notation, we shall represent $\sum_{i=k}^{\ell} C(\bar{p}_1,\ldots,\bar{p}_s,\bar{q}_{s+1},\ldots,\bar{q}_\ell,i)$ by $\sum_{i=k}^{\ell} C(\ell,s,i)$ where s indicates the number of E(p)'s; i.e.,

$$\sum_{i=k}^{\ell} C(\ell,s,i) = \sum_{i=k}^{\ell} C(\underbrace{E(p),\ldots,E(p)}_{s},\underbrace{E(q),\ldots,E(q)}_{\ell-s},i) \qquad (3.2)$$

We compute each $C(\ell,s,i)$, $k\le i\le \ell$, according to the definition of C: since there are exactly i attributes in common between the query and a desired record, there are ℓ-i attributes not in common. j of these ℓ-i attributes ($0\le j\le \ell-i$) can be chosen out of the ℓ-s (1-E(q))'s and the remaining (ℓ-i-j) chosen from the s (1-E(p))'s. Summing j from 0 to ℓ-i,

---

†Owing to the absence of knowledge of the exact values of the p's and the q's, uncertainties are introduced into the expression $mC(p_1,\ldots,p_s,q_{s+1},\ldots,q_\ell,i)$, the expected number of records having i attributes in common with the query. Because of the independence assumption, the expression $m\,E_{(p,q)}\{\sum_{i=k}^{\ell} C(p_1,\ldots,p_s,q_{s+1},\ldots,q_\ell,i)\}$ depends only on the expected values of the p's and the q's (see next paragraph for explanation). Hence, the term "unconditional expected number".

$$C(\ell,s,i) = \sum_{j=0}^{\ell-i} \binom{\ell-s}{j} (1-E(q))^{j} (E(q))^{\ell-s-j} \binom{s}{\ell-i-j} (1-E(p))^{\ell-i-j} \times$$

$$\times (E(p))^{s-(\ell-i-j)} , \qquad (3.3)$$

where some of the terms in the sum may be zero.

Thus we have found a relation between the correlation of a query with the representative of a cluster and the unconditional expected number of records having k or more attributes in common with the query. In the next section, we shall compare the average behavior of a cluster relative to that of another variations of different parameters.

3.4 Analysis

Consider the case where there are only two clusters C_1 and C_2. Let Q be the query submitted and

ℓ = the number of attributes of Q.

The retrieval depends on the threshold k,namely,

k = the minimal number of attributes which a record should possess in common with the query in order to be retrieved.

Suppose now it is found that the representative of C_1,R_1, has fewer attributes in common with Q than that of C_2 (i.e., R_2). The purpose of this analysis is to examine the effect on retrieval performance if cluster C_1 is not retrieved. The main criterion of accepting or rejecting C_1 depends on

W, the ratio of the expected number of desired records in $\mathcal{C}_1$ to that of $\mathcal{C}_2$. If it is sufficiently small, then the amount of desired records in $\mathcal{C}_1$ is likely to be scanty compared with that of $\mathcal{C}_2$. Thus the retrieval of cluster $\mathcal{C}_1$ may not be worthwhile in terms of the time required to examine all the records in $\mathcal{C}_1$.

$\mathcal{C}_1$ and $\mathcal{C}_2$ will hitherto be referred as "average clusters" which are defined in the last section, for clusters with same values for

s = the number of attributes that R_1 has in common with Q, namely, $f(Q,R_1)$,

s+d = the number of attributes that R_2 has in common with Q, namely, $f(Q,R_2)$, where d is an integer >0, and

t = the threshold value for determining whether an attribute should be contained in the representatives.

If N_1 and N_2 are the total number of records contained in $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively, then $W=(N_1/N_2)\sum_{i=k}^{\ell} C(\ell,s,i)/\sum_{i=k}^{\ell} C(\ell,s+d,i)$. However, since N_1 and N_2 are fixed relative to the five parameters N_1/N_2 will be disregarded in this analysis. Thus, the ratio becomes

$$W = \sum_{i=k}^{\ell} C(\ell,s,i) / \sum_{i=k}^{\ell} C(\ell,s+d,i) \quad . \qquad (3.4)$$

With all five parameters given, W can be calculated. How-

ever it is desirable to derive expressions which show how W depends on each parameter while some or all others are fixed. In fact, based on such results, it would be possible to indicate regions of the parameter space in which W takes on a small value and hence the rejection of $\mathcal{C}_1$ is justified. Hopefully, the analysis can also lead to a better understanding of cluster-based retrieval process, resulting in more effective implementation of classification and retrieval algorithms.

To start with the analysis,we express W as

$$W = \prod_{j=1}^{d} \alpha_j \tag{3.5}$$

where

$$\alpha_j = \sum_{i=k}^{\ell} C(\ell,s+j-1,i) / \sum_{i=k}^{\ell} C(\ell,s+j,i) \ . \tag{3.6}$$

The fact that $\alpha_j < 1$ and $\alpha_j < \alpha_{j+1}$, $1 \le j \le d$, will be shown as immediate consequences of Lemmas 3.1 and 3.5 respectively. Thus $(\alpha_1)^d \le W \le (\alpha_d)^d$. As a result, when the difference of the correlations of the query with the two representatives increases, the average number of desired records in one cluster ($\mathcal{C}_1$) relative to that of the other ($\mathcal{C}_2$) decreases rapidly. Let W_o be the value of W such that an ordinary user is likely to consider the percentage of desired records he is likely to find in $\mathcal{C}_1$ to be too low for its retrieval to be worthwhile. If d_o is the value of d which produces W_o, then any value $d \ge d_o$ infers a value $W \le W_o$ and will cause possible rejection of $\mathcal{C}_1$. In case that more than two

clusters are involved, the one which is expected to contain the largest number of desired records will be selected to compare with each of the remaining clusters. The choice of W_o depends on the number of clusters involved, in the sense that the value will be adjusted so as to maintain an adequate number of records to be retrieved. This observation applies throughout this section.

Lemma 3.1: $$\sum_{i=k}^{\ell} C(g_1,\ldots,g_\ell,i) = g_1 C(g_2,\ldots,g_\ell,k-1) + \sum_{j=k}^{\ell-1} C(g_2,\ldots,g_\ell,j)$$

for $k>0$.

Proof: By definition, we have

$$C(g_1,\ldots,g_\ell,i) = \sum_{\binom{\ell}{i}} \left(\prod_{k=1}^{i} g_{j_k}\right)\left(\prod_{k=i+1}^{\ell} (1-g_{j_k})\right) . \quad (3.7)$$

In (3.7), g_1 can appear as a g_{j_k} (i.e.,the probability that the first attribute is in common) or as a $(1-g_{j_k})$ (i.e., the probability that the attribute is not in common). If it appears as a g_{j_k}, then out of the remaining $(\ell-1)$ attributes, we have to choose $(i-1)$ g_{j_k}'s to ensure that there are i attributes in common. On the other hand,if g_1 appears as a $(1-g_{j_k})$, then the i attributes in common are chosen from the remaining $(\ell-1)$ attributes. Thus, if $i\neq 0$ (3.7) is equivalent to

$$\sum_{\substack{\binom{\ell-1}{i-1}\\ j_k\neq 1}} g_1\Big(\prod_{k=1}^{i-1} g_{j_k}\Big)\Big(\prod_{k=i}^{\ell-1}(1-g_{j_k})\Big) + \sum_{\substack{\binom{\ell-1}{i}\\ j_k\neq 1}} (1-g_1)\Big(\prod_{k=1}^{i} g_{j_k}\Big)\Big(\prod_{k=i+1}^{\ell-1}(1-g_{j_k})\Big)$$

$$= g_1 C(g_2,\ldots,g_\ell,i-1) + (1-g_1)C(g_2,\ldots,g_\ell,i)$$

$$= g_1[C(g_2,\ldots,g_\ell,i-1) - C(g_2,\ldots,g_\ell,i)] + C(g_2,\ldots,g_\ell,i) \tag{3.8}$$

When i is summed from k to ℓ-1 in (3.8), the coefficients of g_1 cancel each other except the first one $C(g_1,\ldots,g_\ell,k-1)$ and the last one $C(g_1,\ldots,g_\ell,\ell-1)$. We can then add $C(g_1,\ldots,g_\ell,\ell)=g_1C(g_2,\ldots,g_\ell,\ell-1)$ to this equation, and the desired result follows.

By Lemma 3.1, (3.2) and (3.6), if $k>0$,

$$\alpha_j = \sum_{i=k}^{\ell} C(\ell,s+j-1,i) \Big/ \sum_{i=k}^{\ell} C(\ell,s+j,i)$$

$$= \frac{\left(E(q)C(\ell-1,s+j-1,k-1) + \sum_{i=k}^{\ell-1} C(\ell-1,s+j-1,i)\right)}{\left(E(p)C(\ell-1,s+j-1,k-1) + \sum_{i=k}^{\ell-1} C(\ell-1,s+j-1,i)\right)}$$

$$< 1$$

since $E(q)<E(p)$. This leads to the following lemma:

Lemma 3.2: W decreases as d increases.

Proof: by (3.5) and the fact that each $\alpha_j<1$ for $1\le j\le d$.

We now show that as k increases (i.e., the records retrieved will be more "similar" to the query), W decreases.

This result implies that when closer neighbors are required, proportionally fewer desired records will be expected to appear in $\mathcal{C}_1$ relative to $\mathcal{C}_2$. In other words if fewer records are retrieved, the percentage loss of desired records (if $\mathcal{C}_1$ is not retrieved) will decrease. Let k_o be the value of k producing W_o (its definition appears earlier in this section), then for a user specifying a threshold $k \geq k_o$, it is rather safe not to retrieve $\mathcal{C}_1$.

Lemma 3.3: W decreases as k increases, i.e.,

$$\sum_{i=k}^{\ell} C(\ell,s,i) / \sum_{i=k}^{\ell} C(\ell,s+d,i) > \sum_{i=k'}^{\ell} C(\ell,s,i) / \sum_{i=k'}^{\ell} C(\ell,s+d,i)$$

for $d>0$ and $k'>k$.

Proof: Let $B_i = C(\ell,s,i)$ and $A_i = C(\ell,s+d,i)$. It is sufficient to show

$$\left(\sum_{i=k}^{\ell} B_i\right) / \left(\sum_{i=k}^{\ell} A_i\right) > \left(\sum_{i=k+1}^{\ell} B_i\right) / \left(\sum_{i=k+1}^{\ell} A_i\right) \quad 0 \leq k \leq \ell \qquad (3.9)$$

By simple inequality manipulation, it can be shown that (3.9) reduces to

$$B_k / A_k > \left(\sum_{i=k+1}^{\ell} B_i\right) / \left(\sum_{i=k+1}^{\ell} A_i\right) \qquad 0 \leq k \leq \ell \qquad (3.10)$$

By lemma A4 of Appendix II, the desired result follows immediately.

We now relate W to the other three parameters, i.e., t, s and ℓ. Again, we shall examine the behavior of W, when each parameter increases, keeping all others constant. To

do so, each α_i in (3.5) needs to be expressed in another way. By (3.2) and (3.6),

$$\alpha_i = \frac{\sum_{i=k}^{\ell} C(x,r_1,\ldots,r_{\ell-1},i)}{\sum_{i=k}^{\ell} C(y,r_1,\ldots,r_{\ell-1},i)} \qquad (3.11)$$

Where $x=E(q)$, $y=E(p)$ and r_j is either $E(p)$, $1\le j\le \ell-1$. Thus, α_i can be considered as a function of x, y and the r_j's.

Lemma 3.4: $\partial\alpha_i/\partial r_j > 0$, for $1\le j\le \ell-1$ and $1\le i\le d$.

Proof: By Lemma 3.1, rewrite (3.11) as

$$\alpha_i = \frac{r_1 C(E(q),r_2,\ldots,r_{\ell-1},k-1) + \sum_{i=k}^{\ell-1} C(E(q),r_2,\ldots,r_{\ell-1},i)}{r_1 C(E(p),r_2,\ldots,r_{\ell-1},k-1) + \sum_{i=k}^{\ell-1} C(E(p),r_2,\ldots,r_{\ell-1},i)} .$$

Because of the symmetry of α_i with respect to the r's, it is sufficient to show that $\partial\alpha_i/\partial r_1 > 0$.

$$\frac{\partial\alpha_i}{\partial r_1} = \frac{\left(\sum_{i=k}^{\ell-1} C(E(p),r_2,\ldots,r_{\ell-1},i)\right) C(E(q),r_2,\ldots,r_{\ell-1},k-1) -}{\left(\sum_{i=k}^{\ell} C(E(p),r_1,r_2,\ldots,r_{\ell-1},i)\right)^2}$$

$$- \frac{\left(\sum_{i=k}^{\ell-1} C(E(q),r_2,\ldots,r_{\ell-1},i)\right) C(E(p),r_2,\ldots,r_{\ell-1},k-1)}{}$$

$$\cdots\cdots$$

$$> 0 \qquad (3.12)$$

$$\Longleftrightarrow \frac{C(E(q),r_2,\ldots,r_{\ell-1},k-1)}{C(E(p),r_2,\ldots,r_{\ell-1},k-1)} > \frac{\sum_{i=k}^{\ell-1} C(E(q),r_2,\ldots,r_{\ell-1},i)}{\sum_{i=k}^{\ell-1} C(E(p),r_2,\ldots,r_{\ell-1},i)} \quad . \tag{3.13}$$

But (3.13) is equivalent to (3.10) where d=1 and ℓ and k are decremented by 1.

Lemma 3.5: W increases as s increases.

Proof: When s increases by 1, one of the r's in each α_i of the form (3.11), say r_j, which is E(q), is increased to E(p). By (3.5),

$$\frac{\partial W}{\partial r_j} = \sum_{x=1}^{d} \left(\left(\frac{\partial \alpha_x}{\partial r_j}\right) \prod_{\substack{i=1 \\ i\neq x}}^{d} \alpha_i \right) > 0$$

since by Lemma 3.4, each $\partial\alpha_x/\partial r_j > 0$.

This lemma has the following implications. Suppose Q' is another query which also has ℓ attributes but a lower correlation (i.e., s decreases) with the two representatives than the original query Q. Assume also that the difference in correlations between the two representatives in relation to Q' is the same as that of Q (i.e., d is fixed). Then the proportion of the expected number of desired records in $\mathcal{C}_1$ to that of $\mathcal{C}_2$ with respect to Q' is lower than that of Q. Thus, Q' is expected to have higher performance with regard

to $\mathcal{C}_2$ as compared with $\mathcal{C}_1$ than is Q. The fact that $\alpha_i < \alpha_{i+1}$, $1 \le i \le \ell - s$ also follows immediately from lemma 3.5.

Lemma 3.6: W increases as ℓ increases.

Proof: By Lemma 3.1,

$$\sum_{i=k}^{\ell+1} C(g_1,\ldots,g_{\ell+1},i) = g_{\ell+1} C(g_1,\ldots,g_\ell,k-1) + \sum_{i=k}^{\ell} C(g_1,\ldots,g_\ell,i) .$$

Thus,

$$\sum_{i=k}^{\ell+1} C(g_1,\ldots,g_\ell,0,i) = \sum_{i=k}^{\ell} C(g_1,\ldots,g_\ell,i) .$$

When ℓ increases by 1 (without increasing s), the new ratio

$$\begin{aligned} W_{\ell+1} &= \sum_{i=k}^{\ell+1} C(x,r_1,\ldots,r_{\ell-1},E(q),i) \Big/ \sum_{i=k}^{\ell+1} C(y,r_1,\ldots,r_{\ell-1},E(q),i) \\ &> \sum_{i=k}^{\ell+1} C(x,r_1,\ldots,r_{\ell-1},0,i) \Big/ \sum_{i=k}^{\ell+1} C(y,r_1,\ldots,r_{\ell-1},0,i) \\ &= \sum_{i=k}^{\ell} C(x,r_1,\ldots,r_{\ell-1},i) \Big/ \sum_{i=k}^{\ell} C(y,r_1,\ldots,r_{\ell-1},i) \end{aligned}$$

which is then equal to the old ratio W_ℓ. The inequality holds since by Lemma 3.4, $\partial\alpha_i/\partial r_\ell > 0$ (for r_ℓ increases from 0 to E(q)).

This result can be interpreted as follows. For a query Q" which has more attributes than the original query Q (i.e., ℓ increases) but the correlations of Q" with the representatives are the same as those of Q (i.e., s and d are fixed), the proportion of the number of desired records in $\mathcal{C}_1$ to that of $\mathcal{C}_2$ with respect to Q"is higher than that of Q. Thus, Q" is expected to perform better than Q on $\mathcal{C}_1$

as compared to G_2.

Finally, we deal with the parameter t. As t increases, E(p) and E(q) are expected to increase, which then affect the value of α_i (by (3.11)), and hence W. It will be shown that dW/dt> 0. For clusters employing a higher threshold t in defining their representatives while keeping the other parameters fixed with query Q, the ratio W increases. This means roughly that if the process of choosing attributes in the representative is more selective, W increases. However, there is a limitation: all other parameters must remain unchanged. As t increases for a given set of clusters, some of the p's may be below the new threshold, causing them to become q's. Thus, with respect to a given query, the parameters s and d may be altered; s is expected to decrease, but the behavior of d is unpredictable. Therefore, it is not possible to analyse the behavior of the ratio W when t changes in a given sets of clusters having the same d, s, ℓ and k but different t.

By (3.5),

$$\frac{dW}{dt} = \sum_{i=1}^{d} \left(\prod_{\substack{j=1 \\ j\neq i}}^{d} \alpha_j \right) \frac{d\alpha_i}{dt}$$

Since $\alpha_j > 0$, $1 \leq j \leq d$, it is sufficient to show that $d\alpha_i/dt > 0$.

Lemma 3.7: If $d(E(q))/dt \geq d(E(p))/dt > 0$[†], then $d\alpha_i/dt > 0$.

[†]See Remark 1 right after this lemma for explanation.

Proof:

$$\frac{d\alpha_i}{dt} = \left(\frac{\partial \alpha_i}{\partial (E(q))} \frac{d(E(q))}{dt} + \frac{\partial \alpha_i}{\partial (E(p))} \frac{d(E(p))}{dt}\right) + \left(\sum_{j=1}^{\ell-1} \left(\frac{\partial \alpha_i}{\partial r_j}\right)\left(\frac{dr_j}{dt}\right)\right). \tag{3.14}$$

It is sufficient to show that each of the two terms in (3.14) is larger than zero. By Lemma 3.1, (3.11) can be rewritten as

$$\frac{\left(E(q)C(r_1,\ldots,r_{\ell-1},k-1) + \sum_{i=k}^{\ell-1} C(r_1,\ldots,r_{\ell-1},i)\right)}{\left(E(p)C(r_1,\ldots,r_{\ell-1},k-1) + \sum_{i=k}^{\ell-1} C(r_1,\ldots,r_{\ell-1},i)\right)}$$

Thus,

$$\left(\frac{\partial \alpha_i}{\partial (E(q))} \frac{d(E(q))}{dt}\right) + \left(\frac{\partial \alpha_i}{\partial (E(p))} \frac{d(E(p))}{dt}\right) = C(r_1,\ldots,r_{\ell-1},k-1) \cdot$$

$$\cdot \left\{ \frac{C(r_1,\ldots,r_{\ell-1},k-1)\left(E(p)\frac{d(E(q))}{dt} - E(q)\frac{d(E(p))}{dt}\right)}{\left(E(p)C(r_1,\ldots,r_{\ell-1},k-1) + \sum_{i=k}^{\ell-1} C(r_1,\ldots,r_{\ell-1},i)\right)^2} + \right.$$

$$\left. \frac{+ \sum_{i=k}^{\ell-1} C(r_1,\ldots,r_{\ell-1},i)\left(\frac{d(E(q))}{dt} - \frac{d(E(p))}{dt}\right)}{\ldots\ldots\ldots\ldots} \right\} > 0$$

since $E(p)>E(q)$ and $d(E(q))/dt \geq d(E(p))/dt$. By Lemma 3.4, $\partial \alpha_i/\partial r_j > 0$ for $1 \leq j \leq \ell-1$.

Remark 1: If the distributions of the p's and the q's are such that their means occur in the middle of the ranges, then $E(q)=t/2$ and $E(p)=(1+t)/2$. Thus $d(E(q))/dt=d(E(p))/dt=1/2$ and the hypothesis of the lemma is satisfied. One such distribution is the uniform distribution.

The actual behavior of W will be explored in the next section by assuming a set of values for the five parameters. The choice of these values is dictated by the results obtained in this section.

3.5 Empirical Results

In this section, the values of α_1, as defined in (3.5) and (3.6) of section 3.4 are obtained for various values of ℓ, k, t and s. α_1 is important because $W \geq (\alpha_1)^d$ as shown in section 3.4. As the behavior of the ratio W (and therefore α_1) with respect to the different parameters is such as described in section 3.4, it is sufficient to obtain the values of α_1 at certain discrete points of the parameters. Its values at other points can be interpolated from the given points. E(p) and E(q) are assumed to be (1+t)/2 and t/2 respectively. In Tables 1-4, the values of α_1 and the minimum value of d such that the rejection of the cluster C_1 is likely to be acceptable to a user, are presented for $0.0 \leq t \leq 0.2$, $k/4 \leq s \leq 3k/4$ and $4 \leq \ell \leq 16$. It can be reasonably assumed that a user is likely to consider it to be more advantageous not to retrieve C_1 when the concentration of desired records in C_1 to that of C_2 is equal or less than W_o. Here W_o is arbitrarily assigned a low value of 10%. When s is close to k, quite a few records in C_1 are likely to meet the user's criteria for retrieval. Its rejection will thus lead to unsatisfactory results. Hence,

the values of α_1 are not shown for $s>3k/4$. Similarly, if t is chosen to be greater than 0.2, then attributes which have rather high probability of occurrences in the cluster, though not as high as t, are removed from the representative, and retrieval performance will substantially deteriorate. As a consequence, we examine the case $0.0 \le t \le 0.2$ only. The results are presented as follows.

(i) For retrieving very close neighbors, i.e., $k \ge 3\ell/4$, a small difference in correlations between the query and the representatives, $d \ge 2$, is sufficient to bring W below W_o for $0.0 \le t \le 0.1$, as illustrated by Tables 1,2,3,4 (c). By (3.5) in section 3.4, as d increases linearly, W decreases more or less exponentially. Thus, for $d \ge 2$, rejecting cluster $\mathcal{C}_1$ will result in the loss of very few desired records in comparison to those retrieved in $\mathcal{C}_2$. On the other hand, when the user is not very selective (i.e. k is not large compared with ℓ, such as $k=\ell/4$), the representative R_2 of $\mathcal{C}_2$ has to be much closer to the query than R_1 of $\mathcal{C}_1$ (i.e., d is large) in order that the rejection of $\mathcal{C}_1$ is justified. In some cases (e.g., some entries in Tables 1-4a), even when all the attributes of the query are included in R_2, W is still well above W_o. Under these circumstances, the rejection of $\mathcal{C}_1$ is surely unacceptable.

As predicted by the results of last section, the values of d to bring W down to W_o when $k=\ell/2$ are between those obtained for $k=\ell/4$ and $k=3\ell/4$, as illustrated by the

entries of Tables 1-4(b) compared to 1-4(a) and (c).

(ii) When the representative of $\mathcal{C}_1$ is not too close to the query compared with the threshold value k, e.g., $s \leq k/4$, the majority of the records in $\mathcal{C}_1$ will likely be considered as undersirable by the user. In this case, for medium value of k, e.g., $k \geq \ell/2$, a small value for d, $d \leq 3$ is sufficient to bring W below W_o for $0.0 \leq t \leq 0.1$, as shown in the leftmost columns of Tables 1, 2, 3, 4(b,c). However, as s goes up to $3k/4$, the situation is similar to that when the threshold value k is low, as described in (i). In some cases (e.g., some entries in the rightmost columns of Tables 1-2(a,b)), no value of d can make W sufficiently small.

(iii) If the length of the query increases, it is expected that the user's retrieval criterion will be raised in terms of the threshold value k and the representative will have more attributes in common with the query. Suppose now $k=c_1\ell$, $s=c_2k$ for some constants c_1 and c_2 ($c_1, c_2 \in \{1/4, 1/2, 3/4\}$ respectively in the tables). As ℓ increases, then most of the α_i's (α_i, for $i \geq 2$, are not shown in the tables) remain unchanged or decrease (compare Tables 1-4 with 5). As a consequence, the d's do not have to increase to maintain $W \leq W_o$. The results of (i) and (iii) are therefore applicable to queries having at least four or more attributes i.e., $\ell \geq 4$.

(iv) Let t_1 and t_2 be values of t, $t_1 \neq 0$ and $t_2 \neq 0$, related by $t_1 = ct_2$ for some constant c (in the tables, c=2, 3 and 4 for $t_1=0.05$; and c=3/2, 2 for $t_2=0.1$, etc). Then from

Tables 1-5, $\alpha_{i_1} \leq c\alpha_{i_2}$ where α_{i_1} and α_{i_2} are the values of α_i's corresponding to the thresholds t_1 and t_2 respectively. In other words, the growth rate of α_i with respect to t is less than linear. Thus, if W_i is the ratio corresponding to the threshold t_i, i=1,2, then $W_1 \leq c^d W_2$. This would allow us to estimate the retrieval performance for clusters using some threshold t, which is not tabulated.

Summarizing the results, the following conclusion can be reached. Clustered search yields excellent retrieval performance for on-line search, when a few records are required (i.e., high value for k). As more records are desired, retrieval performance will deteriorate. The concentration of desired records in a cluster whose representative has small correlation with the query ($s \leq k/4$) is low relative to that of another cluster whose representative has a slightly higher correlation with the query. Thus the rejection of the former cluster is acceptable to most users. An approximate estimate of the ratio W is also given when t varies. In view of the results tabulated, $t > 0.2$ is likely to be unacceptable.

3.6 Conclusion

The analytical results presented in this chapter depend on only a few essential assumptions. There are some limitations of the results and they should be carefully interpreted. Nevertheless the model does represent a

general approach to analyse this information retrieval process. Further results can be developed by imposing more dependence relationships among the five parameters. However, the combinatorics involved might prove to be difficult to handle. In this respect, the simulation as described in the last section presents some interesting observations.

CHAPTER 4

COMPARISON OF MODELS

4.1 Distribution of Similarities

Despite of the apparent dissimilarities between the models described in the chapters 2 and 3, their basic principles actually do not differ very much. Most of the assumptions are made for purpose of estimating the distribution of similarities (as defined in chapter 1) between the records and the particular query in question. Why is the distribution of similarities important? Given this distribution and the threshold, the percentage of records that are "desired" (or as in the feedback process, the relevant and irrelevant documents retrieved) can be calculated. This quantity provides the basis on which the retrieval system is evaluated. If an additional process is imposed on the system, this distribution of similarities will be altered accordingly, thereby providing a different measure of system effectiveness. Relevance feedback (RF) is such a process. The distributions of similarities between the query and the relevant as well as the irrelevant documents are assumed to be normal and characterised by the means (μ's) and the standard deviations (σ's). Mathematically, RF is a mapping of these μ's and σ's to μ^*'s and σ^*'s. The threshold T is not related to the distributions of similarities here. (In chapter 2, T is transformed to T*

just to make sure the same proportion of documents are retrieved each time, for the sake of comparison only.) In the process of retrieval in clustered files (RCF), this distribution is not explicitly specified. Instead, it is derived from the assumptions at a lower level - the attribute level. (this is why the model adopted here is call "microscopic" as compared to the one for RF). From the assumption that the occurrence of one attribute is independent of the occurrence of any other attribute, the values $C(g_1,\ldots,g_\ell,i)$, $0\le i\le\ell$, are derived. The distribution of similarities for RCF is composed of these values.

As an analog to the random variables defined for the RF, we shall demonstrate how the similarity in RCF can be defined a random variable. Let us first define an one-dimensional binomial distribution x_i for the ith attribute contained in the query:

$$\text{Prob}(x_i=1) = g_i,$$

$$\text{Prob}(x_i=0) = 1-g_i,$$

recalling that g_i is the probability that a record in the cluster contains the ith attribute. The similarity between a record in the data base and the query is then the random variable X, where $X=\sum_i x_i$. The probability generating function of X (in z) is $\prod_i(1-g_i+g_iz)$. If this function is expanded into a polynomial in z, say $\sum_i a_i z^i$, then it can be easily shown that $a_i=C(g_1,\ldots g_\ell,i)$, $0\le i\le\ell$.

Unlike its counterparts in RF, X does not conform to any well-known distribution. Because of the number of independent variables g's in the C-function, the properties of this distribution are largely unknown. However, studies of the simulation results indicate that it is close to a poisson distribution when the sum of g's is small, and gradually evolves to a normal distribution as this sum increases. It can also be shown that it has either a single maximum or two (equal) maxima in the adjacant positions. In this sense, It behaves very much like a binomial distribution, and indeed, it is one, when all the g's are equal.

4.2 Choice of Model

The microscopic model is often applied to the processes in which an attribute can be distinguished from the other attributes. It is not so obvious why the microscopic model rather than the macroscopic one, is adopted for RCF. There, unlike processes such as indexing the distribution of each attribute is not explicitly involved and it seems that only clusters need to be differentiated. The reason for adapting such a model lies in the very nature of the process: the relationship among the query, the representative and the cluster. The system always "prefers" searching the cluster whose representative is "closer" to the query. In order that the process is worthwhile, the "preferred" cluster should yield on the average more desired records than the non-preferred one. (otherwise, the clusters are not

properly organized, a pathological case which is not considered by the analysis.) The outright assumptions on the distribution of similarities between the records in a cluster and the query (as in RF) are not appropiate here, as the representative would then be left out. Transitivity of some sort in the form of query - representative - cluster has to be established. That is, the correlation of the query and the representative, together with the relationship between the representative and the records in the cluster should provide the distribution of similarities between the query and the records in the cluster. One solution to this problem, as is presented in chapter 3, is to string these three things together by means of the frequencies of occurrences of attributes in the clusters. Only high frequency attributes can be included in the representative. As the representative of the perferred cluster, according to the search algorithm, has more attributes in common with the query, the preferred cluster will contain more attributes which are present in the query and are among those attributes occurring most frequently in the cluster. Therefore, the preferred cluster should contain higher percentage of desired records on the average. Indeed, this argument is true, as shown in section 3.4.

Conceivably, one can analyse RF by means of the microscopic model: building up the distributions of similarities from the assumptions (i) and (ii) in chapter 3. Here, the complexity of mathematics seems to be the decisive factor for choosing the macroscopic model instead, as the next section demonstrates.

4.3 Mathematical Manipulation

Comparing the mathematics used in each model and the results derived, one must admit the mathematics used for the RF is more comprehensible and the results are generally more elegant. In fact, at one point in working with RCF, the mathematics got almost out of hand! The complexity in dealing with RCF are due to the discrete quantities involved and the large number of variables.

To simplify the analysis, the records in the RCF are represented by the binary vectors. Were a real number allowed for each component of the n-tuple record, more variables would be needed to describe the distribution of the weight of each attribute within a record. Besides, some more assumptions would have to be added to the model to describe the distribution of similarities, which could no longer be derived by means of the C-function. Unfortunately, when only 0-1 values are allowed in the record, the similarity function becomes discrete valued, so do other quantities, such as the proportions of the desired records within the two clusters, the ratio of these two quantities (which is W), the threshold k etc. As a result, it is difficult to observe the behavior of a quantity when another integer valued quantity varies. Complicated combinatorics is in use rather than the more powerful calculus. Lemma 3.3 which states W increases as k increases is a good example here. The proof (or disproof) would be easier to come by if $\partial W/\partial k$ were allowed. On the other hand, the optimal values

of α and β (in RF) would never have been obtained if some of quantities involved had not been continuous.

We classify the variables used in the models as internal variables or external (input) variables. By internal variables are meant those which are employed to specify the data base with respect to a submitted query, i.e., the distribution of similarities. In RF, all the 6 pairs of μ's and σ's can be classified as internal variables. Their counterparts in RCF are $g_1,\ldots,g_\ell$. There are ℓ such g's and ℓ is dependent on the query submitted. It is very difficult to tell at this point which set of internal variables can be more easily manipulated. However, for RF, some relationships are assumed among the μ's, although the relative magnitudes of the σ's are largely unknown. Ironically, each of the g's must not depend on any others. Otherwise, the distribution of similarities would be much more complicated than it is now, that is to say, if it could be derived at all!

The external variables are those parameters which can somehow be controlled by the system or the user. In RF, they are α and β. The threshold T is also one of the external variables, but it has no impact on the process because both Q and Q* are required, for the purpose of comparison to retrieve the same amount of documents. However, there are five input parameters in the RCF; i.e., ℓ, s, t, k and d. The ratio W depends on all of them, as the analysis shows. Moreover, some of these parameters are

inter-dependent. As a result, the parameter space is 5-dimensional compared to the 2-dimensional plane in RF. One simply cannot proceed in the same manner as in RF to locate regions that guarantee good results and obtain the optimal values for these parameters. It is suspected that the analysis conducted in chapter 3 has come close to the mathematical limitation imposed by the model.

4.4 Summary

With only two input parameters and continuous distributions of similarities, RF is a simpler process as compared to RCF. Coupled with the fact that more stringent conditions are imposed on the composition of the data base, it perhaps comes as no surprise that the analysis for RF is more successful and the results are more impressive. Of course, the analysis on RCF is not without merit. It has succeeded in building a model, as no one has before, for this complicated process from which meaningful results can be drawn, based on a minimum number of assumptions. It has also devised means by which the user can exercise more control on the retrieval process. Eventually, it can lead to implementation of a practical search strategy so that the user can interact with the system to control the number of clusters to be searched.

REFERENCES

BENTLEY, J.L. Multidimensional binary search trees used for associative searching. Comm. ACM 9: 509-516, 1975.

BOOKSTEIN, A. SWANSON, D.R. A decision theoretic foundation for indexing. J. of ASIS 1: 45-50, 1975.

BOOKSTEIN, A., SWANSON, D.R. Probablistic models for automatic indexing. J. Amer. Ass. for Inform. Sci. 25: 312-318, 1974.

BORODIN, A., KERR, L., LEWIS, F. Query splitting in relevance feedback systems. Rep. No. ISR-14 to the NSF, Dept. of Comput. Sci., Cornell U., Ithaca, N.Y., Oct 1968.

BROOKES, B.C. The measures of information retrieval effectiveness. Proposed by Swets. J. Doc. 24: 41-54, 1968.

BURKHARD, W.A., KELLER, R.M. Some approaches to best-match file searching. Comm. ACM 4: 230-236, 1973.

CRAWFORD, R.G., MELZER, H.Z. The use of relevant documents instead of queries in relevance feedback. Rep. No. ISR-34 to the NSF, Dept. of Comput. Sci., Cornell U., Ithaca, N.Y., Oct 1968.

FELLER, W. An introduction to probability and its applications. New York: Wiley, 1967.

GIULIANO, V.E., JONES, P.E. Study and test of a methodology for laboratory evaluation of message retrieval

systems. Rep. ESD-TR-66-405 to the Electronic Systems Div., Arthur D. Little Co., Aug 1966.

HARTER, S.P. The Cranfield II relevance assessments: a critical evaluation. Libr. Quart. 41: 229-243, 1971.

HEINE, M.H. Design equations for retrieval systems based on the Swets model. J. Amer. Ass. for Inform. Sci. 25: 183-198, 1974.

IDE, E. New experiments in relevance feedback. Rep. No. ISR-14 to the NSF, Dept. of Comput. Sci., Cornell U., Ithaca, N.Y., Oct 1968.

VAN RIJSBERGEN, C.J. Further experiments with hierarchical clustering in document retrieval. Inf. Stor. & Retr. 1: 1-14, 1974.

RIORDAN, J. Combinatorial identities. New York: Wiley, 1968.

RIVEST, R.L. Analysis of associative retrieval algorithms. Ph.D. Thesis, Stanford University, 1973.

ROCCHIO, J.J. Document retrieval systems--optimization and evaluation. Ph.D. Thesis, Harvard U., Cambridge, Mass., Rep. No. ISR-10 to the NSF, Harvard Computation Lab., Mar 1966.

ROCCHIO, J.J., SALTON, G. Information search optimization and iterative retrieval techniques. Proc. AFIPS 1965 FJCC, vol.27, pt.1. New York: Spartan Books, 293-305.

SALTON, G. Automatic information organization and retrieval.

New York: McGraw-Hill, 1968.

SALTON, G. The SMART retrieval system - experiments in automatic text processing. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

SALTON, G., LESK, M.E. Computer evaluation of indexing and text processing. J. ACM 1: 8-35, 1968.

SCHKOLNICK, M. Secondary index optimization. Proc. Intl. Conf. on Management of Data, SIGMOD: 186-192, 1975.

SWANSON, D.R. Some unexplained aspects of the Cranfield tests of indexing performance factors. Libr. Quart. 41: 223-228, 1971.

SWETS, J.A. Information retrieval systems. Science 141: 245-250, 1963.

SWET, J.A. Effectiveness of information retrieval methods. Amer. Doc. 20: 72-89, 1969.

YU, C. A formal construction of term classes, J. ACM 1: 17-37, 1975.

YU, C., SALTON, G. Precision weighting - an effective automatic indexing method. J. ACM 1: 76-88, 1976.

Table 1 a.

	k = ℓ/4 (1)								
	s = k/4 (0)			s = k/2 (1)			s = 3k/4 (1)		
	α_1	d_o†	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$
t=0.00	0.00	1	0.000	0.66	3	0.533	0.66	3	0.533
t=0.05	0.17	4	0.101	0.71	3	0.589	0.71	3	0.589
t=0.10	0.30	4	0.193	0.75	3	0.640	0.75	3	0.640
t=0.15	0.40	4	0.276	0.78	3	0.686	0.78	3	0.686
t=0.20	0.48	4	0.352	0.81	3	0.727	0.81	3	0.727

Table 1 b.

	k = ℓ/2 (2)								
	s = k/4 (1)			s = k/2 (1)			s = 3k/4 (2)		
	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.50	2	0.363
t=0.05	0.13	2	0.071	0.13	2	0.071	0.54	2	0.414
t=0.10	0.23	3	0.107	0.23	3	0.107	0.59	2	0.463
t=0.15	0.31	3	0.160	0.31	3	0.160	0.63	2	0.508
t=0.20	0.38	3	0.211	0.38	3	0.211	0.66	2	0.551

Table 1 c.

	k = 3ℓ/4 (3)								
	s = k/4 (1)			s = k/2 (2)			s = 3k/4 (2)		
	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.07	1	0.070	0.09	1	0.090	0.09	1	0.090
t=0.10	0.13	2	0.021	0.16	2	0.078	0.16	2	0.078
t=0.15	0.18	2	0.042	0.22	2	0.116	0.22	2	0.116
t=0.20	0.23	2	0.066	0.28	2	0.154	0.28	2	0.154

† $d_o = \min\{d \mid \prod_{i=1}^{d} \alpha_i \le W_o \text{ or } d+s = \ell\}$, $W_o = 10\%$.

LENGTH OF QUERY= 8

Table 2 a.

	k = ℓ/4 (2)								
	s = k/4 (1)			s = k/2 (1)			s = 3k/4 (2)		
	α_1	d_o†	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.50	6	0.259
t=0.05	0.26	5	0.099	0.26	5	0.099	0.59	6	0.356
t=0.10	0.42	7	0.189	0.42	7	0.189	0.67	6	0.448
t=0.15	0.53	7	0.284	0.53	7	0.284	0.73	6	0.532
t=0.20	0.61	7	0.376	0.61	7	0.376	0.78	6	0.608

Table 2 b.

	k = ℓ/2 (4)								
	s = k/4 (1)			s = k/2 (2)			s = 3k/4 (3)		
	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.10	2	0.013	0.13	2	0.024	0.18	2	0.080
t=0.10	0.19	2	0.045	0.23	2	0.072	0.31	4	0.076
t=0.15	0.27	2	0.086	0.31	3	0.070	0.40	5	0.105
t=0.20	0.33	3	0.061	0.38	4	0.081	0.47	5	0.152

Table 2 c.

	k = 3ℓ/4 (6)								
	s = k/4 (2)			s = k/2 (3)			s = 3k/4 (5)		
	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.07	1	0.070	0.07	1	0.077	0.12	2	0.038
t=0.10	0.13	2	0.018	0.14	2	0.023	0.20	2	0.077
t=0.15	0.18	2	0.037	0.20	2	0.045	0.27	3	0.063
t=0.20	0.23	2	0.059	0.25	2	0.071	0.33	3	0.090

†$d_o = \min\{d \mid \prod_{i=1}^{d}\alpha_i \le W_o \text{ or } d+s=\ell\}$, $W_o = 10\%$.

Table 3 a.

	k = ℓ/4 (3)								
	s = k/4 (1)			s = k/2 (2)			s = 3k/4 (2)		
	α_1	d_o†	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.22	2	0.071	0.32	5	0.089	0.32	5	0.089
t=0.10	0.37	4	0.092	0.48	10	0.167	0.48	10	0.167
t=0.15	0.49	11	0.132	0.59	10	0.269	0.59	10	0.269
t=0.20	0.58	11	0.218	0.67	10	0.373	0.67	10	0.373

Table 3 b.

	k = ℓ/2 (6)								
	s = k/4 (2)			s = k/2 (3)			s = 3k/4 (5)		
	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.11	2	0.014	0.13	2	0.021	0.22	2	0.084
t=0.10	0.20	2	0.048	0.23	2	0.064	0.34	3	0.094
t=0.15	0.28	2	0.090	0.31	3	0.051	0.43	5	0.086
t=0.20	0.35	3	0.060	0.39		0.087	0.51	7	0.105

Table 3 c.

	k = 3ℓ/4 (9)								
	s = k/4 (2)			s = k/2 (5)			s = 3k/4 (7)		
	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_o$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.06	1	0.067	0.08	1	0.080	0.10	2	0.015
t=0.10	0.12	2	0.016	0.14	2	0.024	0.18	2	0.043
t=0.15	0.17	2	0.033	0.20	2	0.047	0.25	2	0.077
t=0.20	0.22	2	0.053	0.26	2	0.073	0.31	3	0.049

† $d_o = \min\{d \mid \prod_{i=1}^{d} \alpha_i \leq W_o \text{ or } d+s = \ell\}$, $W_o = 10\%$.

Table 4 a.

	k = ℓ/4 (4)								
	s = k/4 (1)			s = k/2 (2)			s = 3k/4 (3)		
	α_1	d_o†	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.20	2	0.054	0.26	2	0.094	0.35	4	0.094
t=0.10	0.36	3	0.080	0.42	5	0.087	0.52	13	0.145
t=0.15	0.47	5	0.096	0.54	14	0.137	0.63	13	0.251
t=0.20	0.57	15	0.132	0.63	14	0.232	0.71	13	0.365

Table 4 b.

	k = ℓ/2 (8)								
	s = k/4 (2)			s = k/2 (4)			s = 3k/4 (6)		
	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.10	2	0.012	0.13	2	0.019	0.18	2	0.045
t=0.10	0.19	2	0.041	0.23	2	0.062	0.30	3	0.051
t=0.15	0.27	2	0.080	0.31	3	0.045	0.39	3	0.097
t=0.20	0.34	3	0.048	0.39	3	0.079	0.47	4	0.099

Table 4 c.

	k = 3ℓ/4 (12)								
	s = k/4 (3)			s = k/2 (6)			s = 3k/4 (9)		
	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o}\alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.06	1	0.067	0.07	1	0.077	0.10	2	0.012
t=0.10	0.12	2	0.016	0.14	2	0.021	0.18	2	0.037
t=0.15	0.18	2	0.033	0.20	2	0.042	0.24	2	0.067
t=0.20	0.23	2	0.054	0.25	2	0.067	0.30	3	0.037

†$d_o = \min\{d \mid \prod_{i=1}^{d}\alpha_i \le W_o \text{ or } d+s = \ell\}$, $W_o = 10\%$.

Table 5 a.

	k = ℓ/4 (10)								
	s = k/4 (3)			s = k/2 (5)			s = 3k/4 (8)		
	α_1	d_o†	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.21	2	0.052	0.26	2	0.078	0.37	3	0.081
t=0.10	0.37	3	0.067	0.43	4	0.057	0.55	6	0.083
t=0.15	0.50	4	0.089	0.56	5	0.099	0.66	32	0.111
t=0.20	0.60	6	0.094	0.65	10	0.099	0.75	32	0.240

Table 5 b.

	k = ℓ/2 (20)								
	s = k/4 (5)			s = k/2 (10)			s = 3k/4 (15)		
	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.10	2	0.011	0.13	2	0.018	0.18	2	0.037
t=0.10	0.19	2	0.039	0.23	2	0.057	0.30	3	0.035
t=0.15	0.27	2	0.077	0.32	3	0.037	0.40	3	0.075
t=0.20	0.34	3	0.043	0.39	3	0.068	0.47	4	0.068

Table 5 c.

	k = 3ℓ/4 (30)								
	s = k/4 (8)			s = k/2 (15)			s = 3k/4 (23)		
	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$	α_1	d_o	$\prod_{i=1}^{d_o} \alpha_i$
t=0.00	0.00	1	0.000	0.00	1	0.000	0.00	1	0.000
t=0.05	0.06	1	0.068	0.07	1	0.077	0.10	2	0.011
t=0.10	0.12	2	0.016	0.14	2	0.021	0.18	2	0.035
t=0.15	0.18	2	0.033	0.20	2	0.041	0.25	2	0.065
t=0.20	0.23	2	0.054	0.25	2	0.066	0.30	2	0.098

† $d_o = \min\{d \mid \prod_{i=1}^{d} \alpha_i \leq W_o \text{ or } d+s = \ell\}$, $W_o = 10\%$.

Appendix I

The proofs of the technical lemmas in section 2.4 are given here.

Lemma 2.5. There exists a constant $c_1>0$ and a constant $t_1>0$ such that, for every $t\geq t_1$ and every $0\leq m\leq 1$, the threshold $T^*(m,t) \leq c_1 t$.

Proof. For any (m,t), either $Q^*\geq Q$ or $Q^*\leq Q$. If $Q^*\geq Q$, we have $(T-\mu_1)/\sigma_1 \geq (T^*(m,t) - \mu_1^*)/\sigma_1^*$, which can be rewritten as, by (2.11), $t\{[(T-\mu_1)/\sigma_1](\sigma_1^2/t^2 + m^2\sigma_3^2/r + \sigma_4^2/s)^{1/2} + (\mu_1/t + m\mu_3 - \mu_4)\} \geq T^*(m,t)$. Letting $t_0 = 1$ and $c_0 = |[(T-\mu_1)/\sigma_1](\sigma_1^2 + \sigma_3^2 + \sigma_4^2)^{1/2}| + (\mu_1+\mu_3)$, it is obvious that, when $t\geq t_0$, $c_0 t\geq T^*(m,t)$ for $0\leq m\leq 1$. If $Q^*\leq Q$, we have $(T-\mu_2)/\sigma_2 \geq (T^*(m,t)-\mu_2^*)/\sigma_2^*$. Proceeding similarly as in the former case, we obtain the following result. There exists a constant c_2 and a constant t_2 such that, for $Q^*\leq Q$, $0\leq m\leq 1$ and $t\geq t_2$, $c_2 t\geq T^*(m,t)$. Taking $c_1 = \max\{c_0,c_2\}$, and $t_1 = \max\{t_0,t_2\}$, the desired result follows.

Lemma 2.6. There exists a constant t_5 such that, for every $m>0$ and for every $t\geq t_5$, $\partial I/\partial t\ (m,t) < 0$.

Proof. We shall first establish the result for the case $1\geq m\geq 0$. Referring to (2.15), it is seen that $\partial I/\partial t(m,t)$ is the product of two factors, with the first factor always positive for $0\leq m\leq 1$. By lemma 2.5, it is found that for $t\geq t_1$ and $0\leq m\leq 1$, the second factor $\leq - d_1 t^3 + d_2 t^2 + d_3 t + d_4$ where d_1, d_2, d_3 and d_4 are some positive constants. Thus

there exists a constant t_4 such that, for $t \geq t_4$ and $0 \leq m \leq 1$, the second factor < 0. Taking $t_5 = \max\{t_1, t_4\}$, we then have $\partial I/\partial t(m,t) > 0$, for t t_5 and $0 \leq m \leq 1$. For the case where $m > 1$, let $n = 1/m < 1$ and proceed as in the last lemma.

Lemma 2.7. $\partial I/\partial\alpha(0,\beta) > 0$ and $\partial I/\partial\beta(\alpha,0) > 0$ when $r,s \neq 0$.

Proof. By differentiating (2.8) with respect to α, the expression $\partial T^*/\partial\alpha$ can be derived. Substituting it into the expression $\partial I/\partial\alpha$ obtained by differentiating $I(\alpha,\beta)$ with respect to α (given in section 2.4), we get the following result.

$$\partial I/\partial\alpha = \{(rk_1k_2G_1G_2)/(\sqrt{2\pi}\ (\sigma_1^*)^3(\sigma_2^*)^3 (k_1G_1/\sigma_1^* + k_2G_2/\sigma_2^*))\} \cdot \{(\sigma_1^*)^2(\sigma_2^*)^2 \cdot (\mu_3-\mu_6) - (\sigma_1^*)^2(T^*-\mu_2^*)\alpha\sigma_6^2 + (T^*-\mu_1^*)(\sigma_2^*)^2\alpha\sigma_3^2\}$$

where $G_i = \exp\left(-1/2\ \left(\frac{T^*(\alpha,\beta)-\mu_i^*}{\sigma_i^*}\right)^2\right)$, $i=1,2$.

It is seen that $\partial I/\partial\alpha > 0$ when $\alpha = 0$ since $\mu_3 > \mu_6$. Similarly, we obtain

$$\partial I/\partial\beta = \{(sk_1k_2G_1G_2)/(\sqrt{2\pi}\ (\sigma_1^*)^3(\sigma_2^*)^3 (k_1G_1/\sigma_1^* + k_2G_2/\sigma_2^*))\} \cdot \{(\sigma_1^*)^2(\sigma_2^*)^2(\mu_5-\mu_4) - (\sigma_1^*)^2(T^*-\mu_2^*)\beta\sigma_5^2 + (T^*-\mu_1^*)(\sigma_2^*)^2\beta\sigma_4^2\}.$$

$\partial I/\partial\beta > 0$ when $\beta = 0$ since $\mu_5 > \mu_4$.

Appendix II

Lemma A1:

$$\sum_{j=0}^{\ell-k}\binom{\ell-s}{j}\binom{\ell-j}{\ell-k-j}z^j = \sum_{j=0}^{\ell-k}\binom{\ell-s}{j}\binom{s}{\ell-k-j}(z+1)^j \quad \text{for } z\geq 0$$

Proof:

$$\sum_{j=0}^{\ell-k}\binom{\ell-s}{j}\binom{\ell-j}{\ell-k-j}z^j = \sum_{j=0}^{\ell-k}\binom{\ell-s}{j}\binom{\ell-j}{\ell-k-j}\left(\sum_{i=0}^{j}\binom{j}{i}(-1)^{j-i}(z+1)^i\right)$$

$$= \sum_{i=0}^{\ell-k}\left(\sum_{j=i}^{\ell-k}\binom{\ell-s}{j}\binom{\ell-j}{\ell-k-j}\binom{j}{i}(-1)^{j-i}\right)(z+1)^i$$

$$= \sum_{i=0}^{\ell-k}\left(\sum_{j=i}^{\ell-k}\binom{\ell-j}{\ell-k-j}\binom{\ell-s}{i}\binom{\ell-s-i}{j-i}(-1)^{j-i}\right)(z+1)^i$$

$$= \sum_{i=0}^{\ell-k}\binom{\ell-s}{i}\left(\sum_{j=i}^{\ell-k}\binom{\ell-j}{\ell-k-j}\binom{\ell-s-i}{j-i}(-1)^{j-i}\right)(z+1)^i$$

$$= \sum_{i=0}^{\ell-k}\binom{\ell-s}{i}\left(\sum_{j=0}^{\ell-k-i}\binom{\ell-i-j}{\ell-k-i-j}\binom{\ell-s-i}{j}(-1)^{j}\right)(z+1)^i$$

$$= \left(\sum_{i=0}^{\ell-k}\binom{\ell-s}{i}\right)\binom{s}{\ell-k-i}(z+1)^i \qquad \text{see [Riordan 1968]}$$

Lemma A2

$$\binom{\ell-s}{j}\binom{\ell-s-v}{h-j} - \binom{\ell-s}{h-j}\binom{\ell-s-v}{j} \geq 0 \quad \text{if } j \geq (h-j)$$

Proof: If $h-j>\ell-s$ or $j>\ell-s-v$, the inequality is trivially true since the first expression is non-negative while the other is zero identically. Thus suppose $\ell-s-v\geq h-j$ and $\ell-s\geq j$, then the left hand side is equal to

$$\frac{(\ell-s)!\,(\ell-s-v)!}{j!\,(h-j)!}\left[\frac{1}{(\ell-s-j)!\,(\ell-s-v-(h-j))!} - \frac{1}{(\ell-s-(h-j))!\,(\ell-s-v-j)!}\right]$$

$$= \frac{(\ell-s)!\,(\ell-s-v)!}{j!\,(h-j)!\,(\ell-s-v-(h-j))!\,(\ell-s-(h-j))!}\left[\frac{(\ell-s-(h-j))!}{(\ell-s-j)!} - \frac{(\ell-s-v-(h-j))!}{(\ell-s-v-j)!}\right]$$

$$= \binom{\ell-s}{h-j}\binom{\ell-s-v}{h-j}\frac{(h-j)!}{j!}[((\ell-s)-(h-j))\ldots.(\ell-s-j+1)-(\ell-s-(h-j)-v)\ldots.$$

$$\ldots.(\ell-s-j-v+1)]$$

$$\geq 0$$

since there are equal number of terms inside the square bracket and each term on the left hand side is larger than the corresponding term on the right hand side. Equality occurs when $h-j>\ell-s-v$ or $j>\ell-s$.

Lemma A3

$$\binom{\ell-j}{\ell-k-j}\binom{\ell-(h-j)}{\ell-k-1-(h-j)} - \binom{\ell-(h-j)}{\ell-k-(h-j)}\binom{\ell-j}{\ell-k-1-j} \geq 0$$

if $j\geq(h-j)$.

Proof: The left side of the inequality

$$= \binom{\ell-j}{\ell-k-j}\binom{\ell-(h-j)}{\ell-k-(h-j)}\frac{\ell-k-(h-j)}{k+1} - \binom{\ell-(h-j)}{\ell-k-(h-j)}\binom{\ell-j}{\ell-k-j}\frac{\ell-k-j}{k+1}$$

$$= \binom{\ell-j}{\ell-k-j}\binom{\ell-(h-j)}{\ell-k-(h-j)}\frac{j-(h-j)}{k+1} \geq 0 .$$

Lemma A4: If $B_i=C(\ell,s,i)$ and $A_i=C(\ell,s+d,i)$, then

$$B_k/A_k > \Big(\sum_{i=k+1}^{\ell} B_i\Big)\Big/\Big(\sum_{i=k+1}^{\ell} A_i\Big) \quad , \quad 0\le k\le \ell$$

Proof: The result can be established if the following inequality is true

$$B_k/A_k > B_{k+1}/A_{k+1} > \dots > B_\ell/A_\ell \quad .$$

It suffices to show

$$B_k/A_k > B_{k+1}/A_{k+1} \quad . \tag{3.15}$$

By (3),

$$B_k = \sum_{j=0}^{\ell-k} \binom{\ell-s}{j}(1-E(q))^j (E(q))^{\ell-s-j} \binom{s}{\ell-k-j}(1-E(p))^{\ell-k-j}(E(p))^{s-(\ell-k-j)}$$

$$= (E(p))^{s-(\ell-k)}(1-E(p))^{\ell-k}(E(q))^{\ell-s} \sum_{j=0}^{\ell-k} \binom{\ell-s}{j}\binom{s}{\ell-k-j}\left(\frac{(1-E(q))E(p)}{(1-E(p))E(q)}\right)^j$$

$$= (E(p))^{s-(\ell-k)}(1-E(p))^{\ell-k}(E(q))^{\ell-s} \sum_{j=0}^{\ell-k} \binom{\ell-s}{j}\binom{\ell-j}{\ell-k-j} z^j \quad ,$$

by Lemma A1 where

$$z = [E(p)(1-E(q))]/[E(q)(1-E(p))] - 1.$$

It is easily seen that

$$z = [E(p)-E(q)]/[E(q)(1-E(p))] > 0 \quad ,$$

as required by Lemma A1. After cancelling the terms in common, (3.15) is equivalent to

$$\left(\sum_{j=0}^{\ell-k}\binom{\ell-s}{j}\binom{\ell-j}{\ell-k-j}z^j\right)\left(\sum_{i=0}^{\ell-k-1}\binom{\ell-s-d}{i}\binom{\ell-i}{\ell-k-1-i}z^i\right)$$

$$> \left(\sum_{j=0}^{\ell-k}\binom{\ell-s-d}{j}\binom{\ell-j}{\ell-k-j}z^j\right)\left(\sum_{i=0}^{\ell-k-1}\binom{\ell-s}{i}\binom{\ell-i}{\ell-k-1-i}z^i\right). \quad (3.16)$$

Grouping the terms in ascending powers of z, (3.16) is equivalent to

$$\sum_{h=0}^{2\ell-2k-1}\left(\sum_{j=0}^{\ell-k}a_j b_{h-j}\right)z^h > \sum_{h=0}^{2\ell-2k-1}\left(\sum_{j=0}^{\ell-k}a'_j b'_{h-j}\right)z^h \quad (3.17)$$

where

$$a_j = \binom{\ell-s}{j}\binom{\ell-j}{\ell-k-j}, \quad a'_j = \binom{\ell-s-d}{j}\binom{\ell-j}{\ell-k-j}, \quad 0\le j\le \ell-k$$

$$b_i = \binom{\ell-s-d}{i}\binom{\ell-i}{\ell-k-1-i}, \quad b'_i = \binom{\ell-s}{i}\binom{\ell-i}{\ell-k-1-i}, \quad 0\le i\le \ell-k-1$$

and other a_j, a'_j, b_i and $b'_i = 0$.

(3.17) can be established by showing

$$\sum_{j=0}^{\ell-k} a_j b_{h-j} > \sum_{j-0}^{\ell-k} a'_j b'_{h-j}, \quad 0\le h\le 2\ell-2k-1 \quad (3.18)$$

It is, therefore, sufficient to show[†]

$$a_j b_{h-j} + a_{h-j} b_j \ge a'_j b'_{h-j} + a'_{h-j} b'_j, \quad 0\le j\le \ell-k \quad (3.19)$$

[†]It is easily seen from Lemmas A2 and A3 of Appendix 1 that strict inequality is obtained for some values of j in (3.19) Thus, strict inequality is also obtained in (3.18).

Substituting the values of a_j, a'_j, b_i and b'_i back into (2.19), the following inequality is obtained [††]

$$\left(\binom{\ell-s}{j}\binom{\ell-s-d}{h-j}-\binom{\ell-s}{h-j}\binom{\ell-s-d}{j}\right)\left(\binom{\ell-j}{\ell-k-j}\binom{\ell-(h-j)}{\ell-k-1-(h-j)}-\right.$$

$$\left.-\binom{\ell-(h-j)}{\ell-k-(h-j)}\binom{\ell-j}{\ell-k-1-j}\right) \geq 0 \qquad (3.20)$$

Because of the symmetry involved in (3.20) between j and (h-j), it is sufficient to show the case where $j \geq (h-j)$. Lemmas A2 and A3 establish the desired result.

[††] Inequality (3.20) is not equivalent to (3.19) when some of the a's or b's = 0. Under this situation, it can be proved by more elaborate procedures that (3.19) holds. For ease of presentation, the above approach is adopted.

Appendix III

Summary of Symbols (for Chapter 3 only)

Q: the query

ℓ: length of the query, i.e., the number of attributes in Q

$\mathcal{C}$: the cluster $\mathcal{C}$

R: Representative of the cluster $\mathcal{C}$

k: the minimal number of attributes which a record should possess in common with the query in order to be retrieved

s: the number of attributes that R_1 has in common with Q

d: the difference between the representatives R_1 and R_2 in correlations with Q

t: the threshold (real) value for determining whether an attribute should be contained in a representative

W: the concentration of desired records in $\mathcal{C}_1$ compared to that of $\mathcal{C}_2$, i.e., the ratio of the (average expected) number of desired records in $\mathcal{C}_1$ to that in $\mathcal{C}_2$

W_o: the threshold value of W such that the user may consider it to be more advantageous not to retrieve $\mathcal{C}_1$ if W is less than or equal to this value

g_i: the probability of occurrence of the ith attribute of Q in the records of a cluster $\mathcal{C}$

p_i: the value of g_i if the ith attribute appears in the representative R, i.e., $g_i \geq t$

q_i: the value of g_i if the ith attribute does not appear in the representative R, i.e., $g_i < t$

E(p): expected value of p

E(q): expected value of q

r_i: $E(p)$ or $E(q)$

$C(g_1,\ldots g_\ell,i)$: the probability that a record in C has exactly i attributes in common with Q

$C(\ell,s,i)$: $C(\bar{p}_1,\ldots,\bar{p}_s,\bar{q}_{s+1},\ldots,\bar{q}_\ell,i)$ where $\bar{p}_i=E(p)$ and $q_i=E(q)$